/OLUME 107 PART B NUMBER 35 SEPTEMBER 1960

The Proceedings

OF

THE INSTITUTION OF ELECTRICAL ENGINEERS

FOUNDED 1871: INCORPORATED BY ROYAL CHARTER 1921

PART B

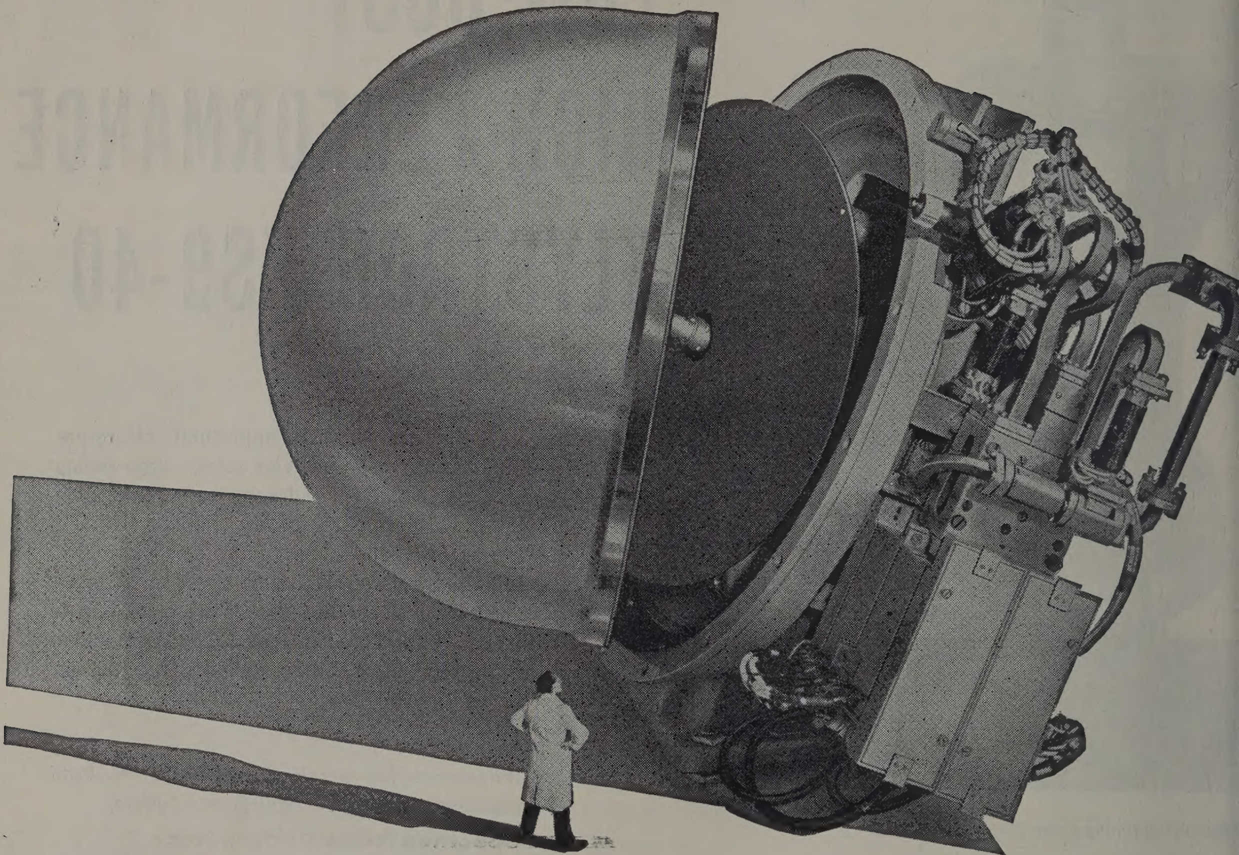
ELECTRONIC AND COMMUNICATION ENGINEERING
(INCLUDING RADIO ENGINEERING)

SAVOY PLACE · LONDON W.C.2

Price Ten Shillings and Sixpence

Where great things are done

with Microwaves



RADAR: Fire Control · Navigation of Aircraft and Small Ships · Automatic Landing · Missile Guidance · Transponders • **COMMUNICATIONS:** Multichannel Radio Links for telemetering Data and Speech • **VALVES:** Klystrons and Magnetrons for 35/GCS and 75/GCS bands · Monitor Diodes for 1/GCS to 35/GCS • **INSTRUMENTS:** Comprehensive Waveguide measuring circuits covering 6 to 75/GCS • **RESEARCH:** Outstanding Research and Development of the latest techniques.

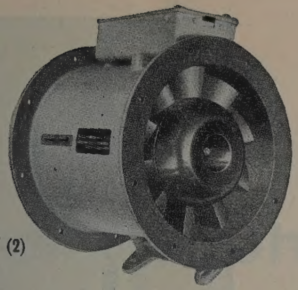
ELLIOTT

COMMUNICATIONS DIVISION · RADAR DIVISION · VALVE DIVISION
MICROWAVE & ELECTRONIC INSTRUMENTS DIVISION · RADAR RESEARCH LABORATORY

ELLIOTT BROTHERS (LONDON) LTD

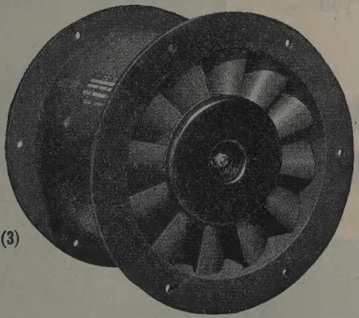
ELSTREE WAY, BOREHAMWOOD, HERTFORDSHIRE · ELSTREE 2040
AIRPORT WORKS, ROCHESTER, KENT · CHATHAM 4/4400

EA *A MEMBER OF THE ELLIOTT-AUTOMATION GROUP*

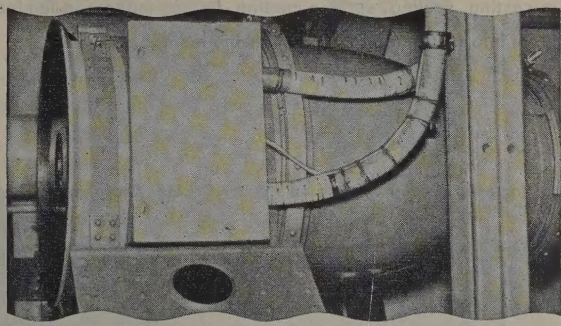


(2)

the top echelon

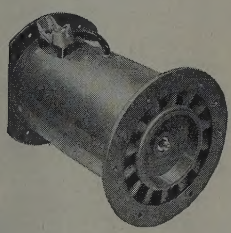


(3)



(1)

# of blowers



(4)

(1) Cabin Air Re-circulating Blower for Vickers Vanguard Aircraft. 9PL241-274 — Single Stage Blower fitted with specially designed self-cooled motor operating on 27V D.C. at a nominal speed of 7,800 r.p.m. Output 1200 c.f.m. at 15.25″ s.w.g. Current input to motor 170 amps. Overall length of unit including tail fairing 16½″. Diameter of Impeller 9.2″. Weight 45 lb.

(2) Standard 6PL121-141 Single Stage Blower. 230V — 1 ph. 50 cs. Performance 280 c.f.m. at ½″ s.w.g. Speed 2,750 r.p.m. Power input 75 watts. Weight 11 lb. 9 oz. Length 6½″. Diameter of Impeller 6″.

(3) 5PL121-122 Single Stage Blower. 230V — 1 ph. 50 cs. Performance 175 c.f.m. at 0.4″ s.w.g. 50 watts input. Speed 2,700 r.p.m. Weight 6 lb. 9 oz. Length 5¼″. Diameter of Impeller 5½″.

(4) 4PL142-138 Two Stage Blower. 115/230V — 1 ph. 50 cs. Performance — Output 60 c.f.m. at 0.4″. Speed 2,720 r.p.m. Power input 90 watts. Weight 7 lb. 4 oz. Overall length 7¾″. Diameter of Impellers 4.3″. Used on ship-borne radio equipment.

Light, compact, finely balanced and ruggedly built, Plannair Blowers have a weight/output ratio that makes them the most efficient in the world.

In range, these Blowers and Wafters include models that will meet the most uncommon requirements of any type of installation where temperature or ventilation is to be regulated by the movement of air.

Write for further details to:

PLANNAIR

PLANNAIR LIMITED
WINDFIELD HOUSE · LEATHERHEAD · SURREY
TELEPHONE: LEATHERHEAD 4091/3 & 2231

PLA46

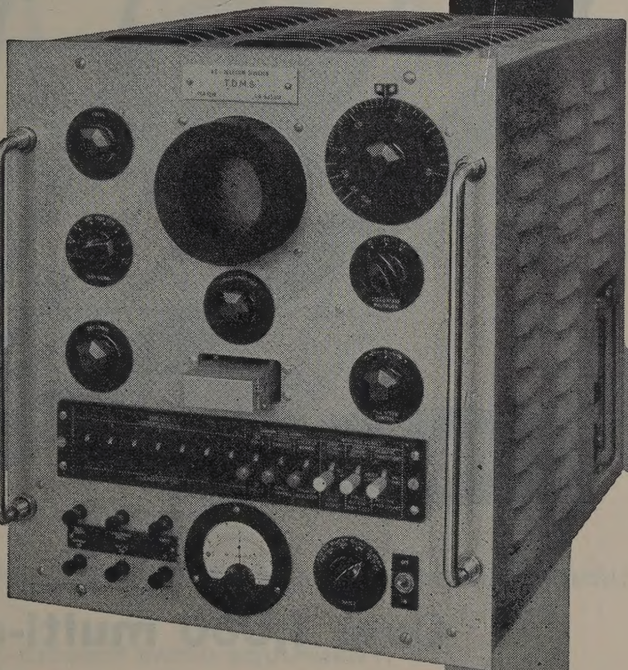
1:1
2:2
6:1
7:7
Q9S
or any single character

Telegraph distortion measurement

in the speed range 40-400 bauds

Rapidly increasing international use of high-speed synchronous telegraph systems has led to the urgent need for suitable distortion measuring equipment. M.S. Type TCA1219, manufactured by A.E.I. at their Woolwich factory to a British Post Office design, enables distortion measurements to be made within 1% accuracy. The method of indication is by bright spots on the circular time base of a cathode-ray tube. Screening and the fitting of radio interference suppressors have been given particular attention.

Size	17 in. high x 15 in. wide x 18 in. deep
Weight	110 lb.
Power supply	200/250 v. 50 c.p.s.
Power consumption	170 w.



ast experience  in the science of telecommunications

AEI **Telecommunications Division**
Transmission Department Woolwich London SE18
Associated Electrical Industries Limited

TA 3775

nts

V.H.F. radio equipments now in production include:

5-CIRCUIT RADIO EQUIPMENT.

Provides five speech circuits, each transmitting the band 300c/s to 3400c/s over distances of up to about 40 miles.

9-CIRCUIT RADIO EQUIPMENT.

Provides nine speech circuits, each transmitting the band 300c/s to 3400c/s with out-of-band signalling, over distances of up to about 40 miles.

TRANSPORTABLE 5-CIRCUIT RADIO EQUIPMENT.

Ideal for temporary or emergency links.

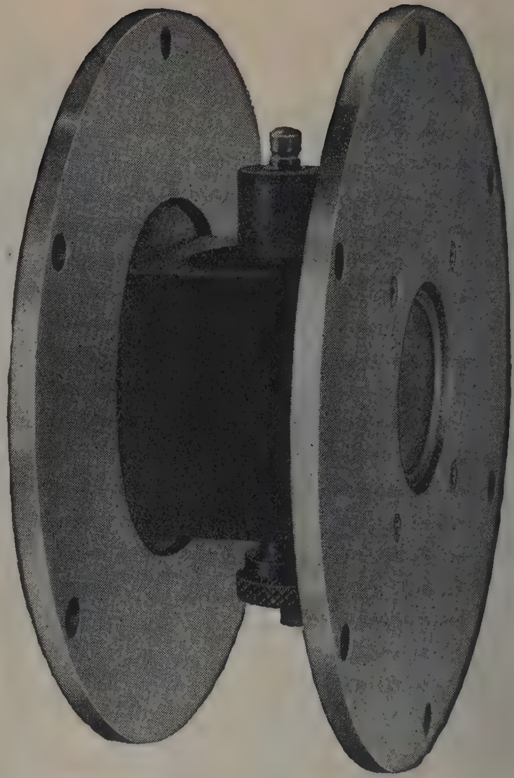
For further details regarding these equipments, please write for standard specifications SPO.5051, SPO.5060 and SPO.5070 respectively.

IS

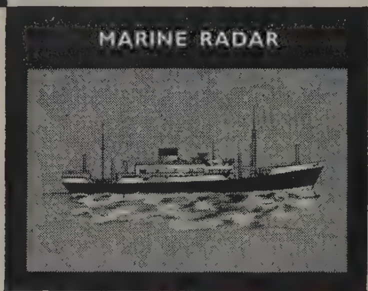
COVENTRY, ENGLAND

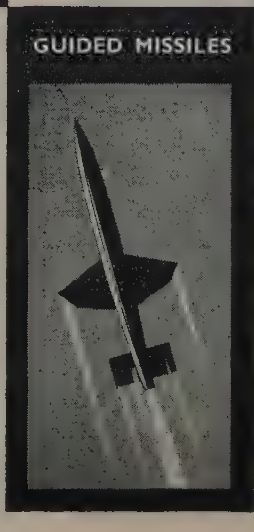
G.E.C. 34

For
Ground, Sea and Air
FERRANTI
T.R.CELLS



Series QF.32
Medium Q S-Band Tunable T.R. Cell
Peak Power: 100KW.
Frequency Range: 2600-3950 M/cs.





A comprehensive range of Ferranti T.R. Cells is available covering frequencies from 1000 Mc/s. to 35000 Mc/s. For further information write to

FERRANTI LTD · KINGS CROSS ROAD · DUNDEE
Telephone: DUNDEE 87141

Ferranti

DS/T68

AEI achievement in cordless switchboards

1934	*The first cordless trunk switchboard in the world installed in Cape Town, South Africa.*
1939	*The first cordless switchboard installed in Australia at Melbourne.*
1950	*The first postwar modernised cordless switchboard in Australia, installed at Adelaide.*
1954	*The first cordless switchboard installed for the British Post Office at Thanet.*
1957	*The largest cordless switchboard in the world installed in Sydney, Australia.*
1959	*The second cordless switchboard installed for the British Post Office at Middlesbrough.*

A.E.I. Telecommunications Division has:

The greatest experience in cordless switchboard design and manufacture.

Installed the largest number of cordless switchboards—over 1,300 positions.

Been the only supplier of cordless switchboards to the majority of Commonwealth users.

Employed its vast worldwide experience in the design of cordless switchboards for the most modern British PABX equipment.

PABX equipments are an extension of the automatic public telephone service into the premises of the subscriber. It is therefore natural that the A.E.I. Telecommunications Division should base its cordless switchboard PABX design upon its worldwide public service experience.

AEI

Associated Electrical Industries Limited
Telecommunications Division
Formerly the Telecommunications Division of Siemens Edison Swan Ltd
Woolwich London S.E.18 WOOlwich 2020

ERICSSON MULTI-CHANNEL TELEGRAPH EQUIPMENT

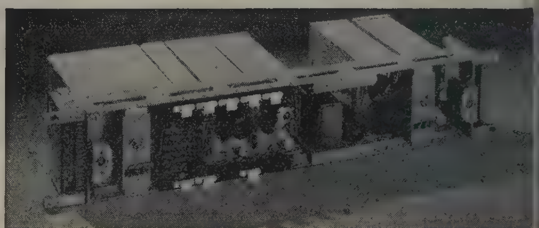
FMT. 241

This FMT. 241 equipment provides 24 frequency modulate voice frequency telegraph channels to C.C.I.T.T. recom mendations over any bearer having a bandwidth of 300-3400 c/s It has considerable advantage over the equivalent amplitud modulated system under conditions of high noise and has wider level range and lower distortion at higher signallin speeds. It is designed to work directly from double curren (two path polar) telegraph signals but can be supplied wit converters (loop control equipment) for use with singl current (neutral) telegraph signals. Conversion to party lin working is easy and suitable conversion equipment can b supplied for specific requirements.

The FMT. 241 equipment is mounted on plug-in panels carrie on two 9 ft. racks equipped with full test facilities or two 6 ft racks with reduced test facilities. Operation is from 90-260 V 40-60 c/s. a.c. mains supply. Auxiliary equipment comprises wall mounted 80-0-80 volt signalling supply furnishing up t 1.5 amps or alternatively a rack mounted supply of 700 mA capacity.

A system using the FMT.241 equipment and givin party line facilities was recently developed an installed for International Aeradio (Caribbean) Ltd This is the largest v.h.f. inter-island aeronautica system in the world, linking ten of the Caribbea group of islands from St. Kitts in the north t Trinidad in the south. Using only one telephon channel as a bearer circuit it serves some 25 tele graph offices or about 60 teleprinter installations

VF Band	360-3240 c/s.
Basic channels	1140-2460 c/s.
Maximum Telegraph Speed	80 bauds
Telegraph Distortion (50 bauds)	3% (nominal)
Power Consumption	380 watts approx.
Rack Dimensions	2 bays 6′ 0″ x 1′ 8½″ x 8″
	or 2 bays 9′ 0″ x 1′ 8½″ x 8″



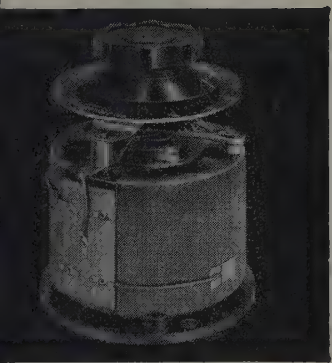
For further information please write to:

ERICSSON TELEPHONES LIMITED — ETELCO LIMITED

HEAD OFFICE:—22 LINCOLNS INN FIELDS, LONDON, W.C.2. TEL: HOL. 6936
WORKS:—BEESTON, NOTTINGHAM AND SUNDERLAND

VOLTAGE CONTROL

A range of A.C. Mains Voltage Stabilisers which incorporate electronic detector systems of new design. They are robust, reliable and very compact.
The unit shown has a rating of 2.5 KVA, will handle input variations from 200-250 Volts whilst maintaining a pre-set controlled voltage between 220-240 Volts to an accuracy of $\pm$ 0.5%
This particular Stabiliser is suitable for bench, floor or wall mounting and is fitted into a steel case attractively finished in hammered aluminium. Dimensions are : Height 15", Width 10", Depth 7½".



Torovolts are the new toroidally wound continuously variable auto-transformers. Electrical and mechanical improvements have been incorporated to give greater safety and reliability. Legible dials ensure maximum convenience in use.

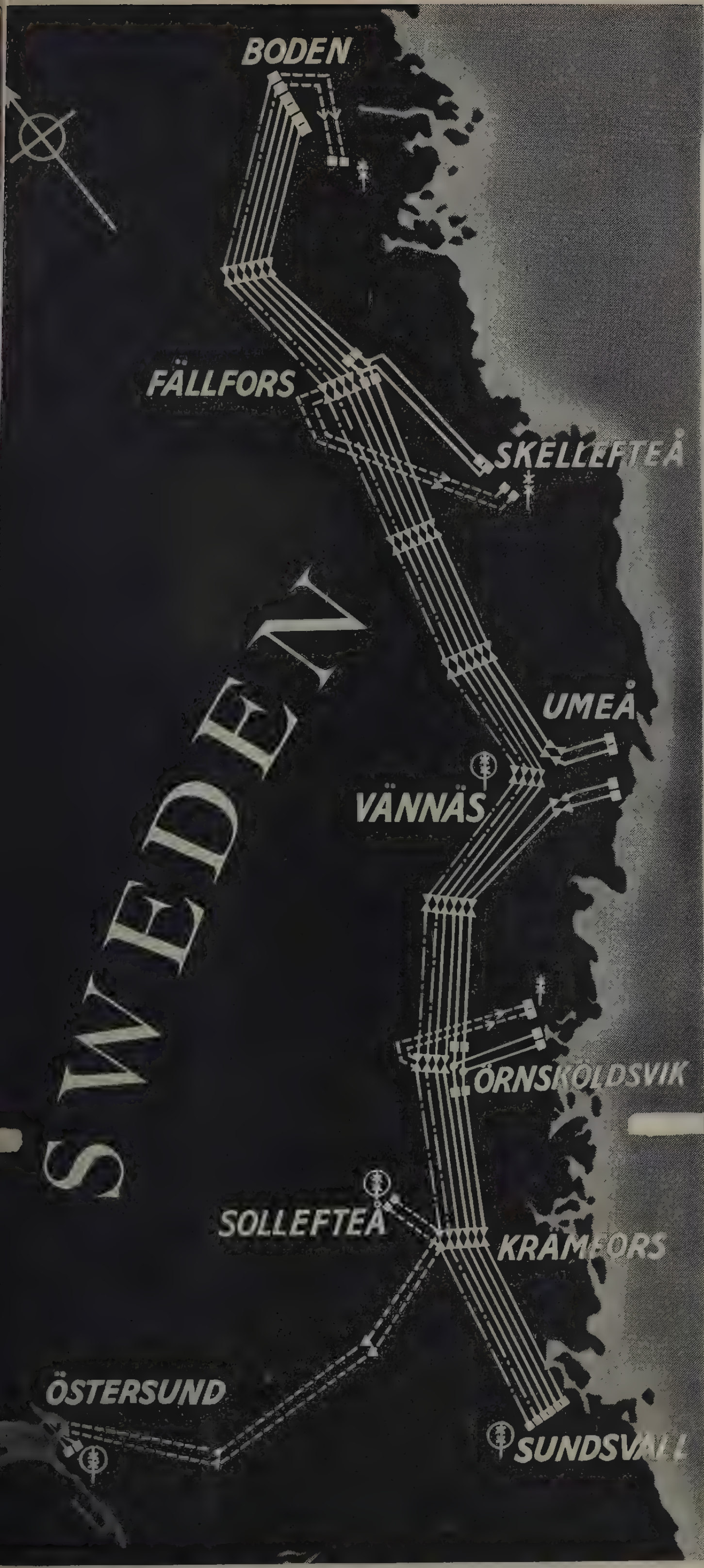
Expamet-Cressall resistors, rheostats and heaters form the most comprehensive range available from a single source. Units to control loads from 4 watts to over 20,000 Kilowatts are already in use all over the world. To widen still further the enormous field of applications of Expamet-Cressall products, two new designs have been introduced and are briefly described here. Further details and technical advice from our engineers is freely available.

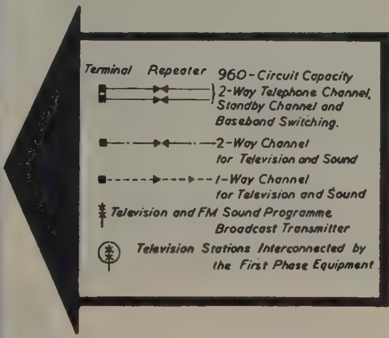
EXPAMET—CRESSALL

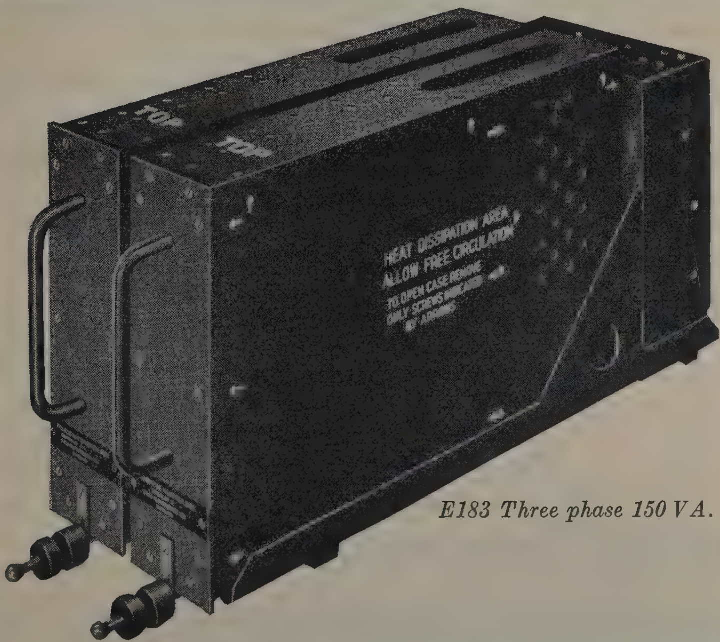
Expamet Cressall

The Electrical Division of
The Expanded Metal Company Ltd.

London Office: 16 Caxton Street, London, S.W.1. Telephone: ABBey 7766
Works: Stranton Works, West Hartlepool. Tel: Hartlepools 5531
The Cressall Manufacturing Company Ltd., Eclipse Works, Tower Street, Birmingham 19
Tel: Aston Cross 2666


BODEN
FÄLLFORS
SKELLEFTEÅ
SWEDEN
UMEÅ
VÄNNÄS
ÖRNSKÖLDSVIK
SOLLEFTEÅ
KRAMFORS
ÖSTERSUND
SUNDSVALL


Terminal
Repeater
960-Circuit Capacity
2-Way Telephone Channel, Standby Channel and Baseband Switching.
2-Way Channel for Television and Sound
1-Way Channel for Television and Sound
Television and FM Sound Programme Broadcast Transmitter
Television Stations Interconnected by the First Phase Equipment



E183 Three phase 150 V A.

THE PROCEEDINGS OF THE
INSTITUTION OF ELECTRICAL ENGINEERS

TEN YEAR INDEX

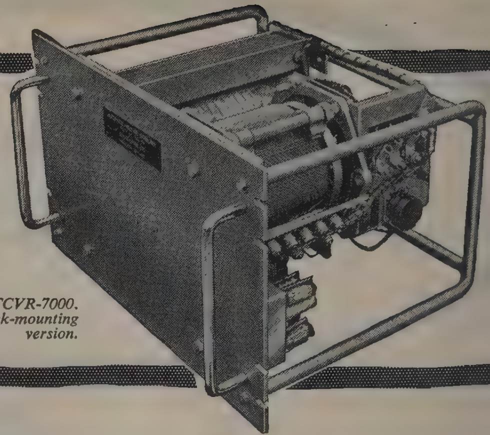
1942—1951

A TEN-YEAR INDEX to the *Journal of The Institution of Electrical Engineers* for the years 1942–48 and the *Proceedings* 1949–51 (vol. 89–98) can be obtained on application to the Secretary.

The published price is £1 5s. 0d. (post free), but any member of The Institution may have a copy at the reduced price of £1 (post free).

40 VOLTS/SEC AUTOMATIC CORRECTION

—with the type TCVR voltage regulator



Type TCVR-7000. Rack-mounting version.

The TCVR is a servomechanical automatic voltage regulator having the very high speed of correction of FORTY VOLTS PER SECOND. It provides an *undistorted* output, maintained constant within very close limits (normally ±0.5%) from no-load to full-load, for wide variations in frequency and power factor.

A wide range of models from 1.6 to 12 kVA single-phase, and 4.8 to 36 kVA three-phase, is available, to standard or tropical specification, in cabinets or for rack-mounting. Models are also available in which the output voltage is continuously adjustable over a wide range by means of a panel control. Regulators can be supplied to Services' specification, and special models can be designed to order.

For high-speed, accurate stabilisation *without distortion*—specify TCVR.

Claude Lyons Ltd.

Other products of Claude Lyons Ltd. Stabiliser Division

BMVR: Motor-driven laboratory and industrial regulators ranging from 1·6 to 29 kVA single-phase, and 4·8 to 87 kVA three-phase. Constancy of output normally ± 0·5%, from no-load to full-load. No distortion. Speed of correction 1 Volt/Sec. A great variety of models, standard, tropical and militarised, for all applications.

BAVR: Electronic stabilisers of very high accuracy, and very rapid response, with no moving parts. Input range : − 10% to + 5%, output constancy ± 0.15%. Three sizes : 200, 500 & 1000 VA. Exceptionally useful for control of chemical processes, heating, lighting, etc.

ASR: Automatic step regulators, small, inexpensive, and with sinusoidal output waveform. Two sizes : 1·15 kVA and 2·3 kVA. Input range − 10% to + 5%; output constancy, ±2½%.

ATC: Automatic Tap-Changing Transformers — a development of ASR. Two sizes : 575 VA and 1150 VA. Input range − 20% to + 10% : output constancy, ± 5%. Provide adequate stabilisation for many types of apparatus, at low cost. Also useful as pre-regulators, e.g. in conjunction with BAVR.

We shall be pleased to send you full details of our entire range.

Stabiliser Division **VALLEY WORKS · HODDESDON · HERTS · TELEPHONE HODdesdon 4541—**

CL/47/E2

THE JOURNAL OF

The British Nuclear Energy Conference

The Institution of Civil Engineers
The Institution of Mechanical Engineers
The Institution of Electrical Engineers
The Institute of Physics
The Institution of Chemical Engineers
The Institute of Metals
The Iron and Steel Institute
The Institute of Fuel
The Joint Panel on Nuclear Marine Propulsion

PUBLISHED JANUARY, APRIL, JULY, OCTOBER

The Journal contains papers and discussions on the applications of nuclear energy and ancillary subjects

ANNUAL SUBSCRIPTIONS:

MEMBERS 30/– post free

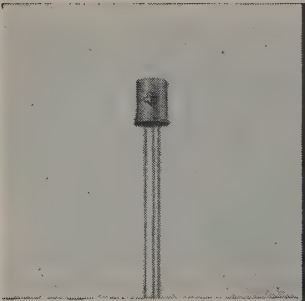
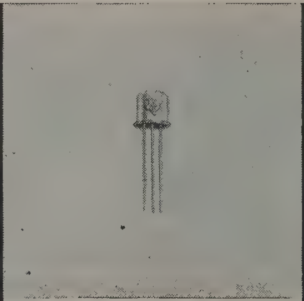
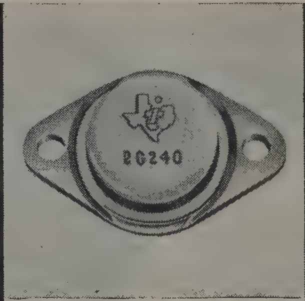
NON-MEMBERS 60/– post free

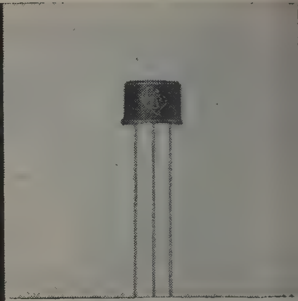
Full particulars are available from

The Secretary • B.N.E.C. • 1-7 Great George Street • London • SW1

INDEX OF ADVERTISERS

COMPUTER TRANSISTORS FROM TEXAS



FOR MEDIUM SPEED COMPUTERS	FOR FAST COMPUTERS	FOR CORE DRIVING	FOR CORE DRIVING
(up to 1 Mc/s digit rates)	(up to 50 Mc/s digit rates)		
2G301 • 2G302	2N711	2G240	2N696 • 2N697
2G303 • 2G304	2G103 • 2G104	I_C=3A V_{CB}=80V	f_1=40 Mc/s (min)
High Frequency • High Current	50mA. 300 Mc/s	f_1=15 Mc/s (min)	I_C=500 mA
GERMANIUM	GERMANIUM	GERMANIUM	SILICON
PNP	PNP	PNP	NPN
ALLOY JUNCTION	DIFFUSED 'MESA'	ALLOY DIFFUSED	DIFFUSED 'MESA'

FOR ALL COMPUTERS - TEXAS

Write for Data Sheets & Application Information on these and other Texas semiconductor devices which include

NPN Silicon Transistors (Signal · UHF · Power)
Germanium Transistors (UHF · Power)
Silicon Rectifiers & Diodes
(High Voltage · Signal · Computer · Photo · Zener)

Silicon Zener Voltage Regulators
Silicon Controlled Rectifiers
Semiconductor-Grade Silicon

TEXAS INSTRUMENTS LIMITED

DALLAS ROAD BEDFORD ENGLAND
BEDFORD 68051 CABLES: TEXINLIM · BEDFORD

91/A26

The Institution is not, as a body, responsible for the opinions expressed by individual authors or speakers.
An example of the preferred form of bibliographical references will be found beneath the list of contents.

THE PROCEEDINGS OF THE INSTITUTION OF ELECTRICAL ENGINEERS

EDITED UNDER THE SUPERINTENDENCE OF W. K. BRASHER, C.B.E., M.A., M.I.E.E., SECRETARY

VOL. 107. PART B. No. 35. SEPTEMBER 1960

537.59

The Institution of Electrical Engineers
Paper No. 3301
Sept. 1960
©

THE FIFTY-FIRST KELVIN LECTURE

'COSMIC RADIATION'

By Professor C. F. POWELL, M.A., Ph.D., Sc.D., F.R.S.

(*Lecture delivered before* THE INSTITUTION, 31*st March*, 1960.)

Coming out of space, and incident on the high atmosphere, there is a thin rain of energetic charged particles known as the primary cosmic radiation. The work of the past twenty years has shown that most of the incoming particles of great energy are atomic nuclei stripped of all the electrons with which they are normally surrounded in ordinary matter. These particles approach the solar system equally from all directions; their directions of motion are distributed isotropically.

The existence of cosmic radiation poses a number of problems of great contemporary interest in the general field of cosmology. Where do the particles originate, and by what physical processes are they endowed with their great energy? Do they come from the sun and similar stars? Are they confined to the galactic systems, or do they pervade the space between them? These are some of the problems which have led to an intensive study of cosmic radiation in the past twenty years. The possibility of making a serious approach to these problems owes a great deal to recent advances in our understanding of the evolution of stars and of the way the heavier elements of the periodic classification are built up from the primeval hydrogen at high temperatures in stellar interiors. A second important new source of knowledge closely related to our problems is radio-astronomy, which has given us a new insight into the structure of our own star-city —our galaxy—and the conditions in the diffuse halo of gas, nearly spherical in form, with which it is now known to be surrounded.

The Nature of the Cosmic Radiation

Before attempting to speculate about the origin of cosmic radiation, it was necessary to establish in detail the characteristics of the radiation arriving at the top of the atmosphere. I have spoken of the radiation as being composed of atomic nuclei. But what nuclei? And in what proportions? We speak of this as the problem of the 'charge spectrum' of the cosmic radiation. And we need to know in more detail how many particles there are with different values of the energy; we speak of this as the 'energy spectrum' of the radiation.

Prof. Powell is Melville Wills Professor of Physics, University of Bristol.

We have known since 1948, as a result of experiments by Freier and others with balloons, that in addition to hydrogen nuclei, which are the most numerous of the incoming particles in the primary cosmic radiation, there are also present those of heavier elements. Elements up to numbers 26 or 28 in the periodic classification—nickel and iron—have been detected. Fig. 1 shows the tracks of three nuclei of the cosmic radiation as recorded in photographic emulsions carried to great altitudes by balloon. A photographic emulsion, of the type employed in nuclear physics, consists of myriads of small crystals of silver halide—mostly the bromide—suspended in gelatine. When a charged particle passes through such an emulsion it changes slightly some of the grains of the silver halide which it traverses, so that when the plate is developed they are reduced to black grains of silver. The paths of charged particles may thus be made manifest. After fixation, washing and drying, the resulting tracks may be examined under the microscope and photographs of them taken. The pictures are made from a succession of photomicrographs of the tracks in the emulsion. In modern conditions, large stacks composed of hundreds of sheets of emulsion, each about 0·5 mm thick, are assembled together to form a solid sensitive mass. After exposure, the individual sheets are dipped in a gelatine solution and rolled on to specially treated glass plates to which they are thus made to adhere. They are then processed by methods similar in principle to those employed in conventional photography. The largest stack employed, hitherto, had a volume of 80 litres and weighed about 700 lb. The processing of such a stack is a protracted operation involving the use of about 20 tons of hypo in the fixing baths.

From Fig. 1 it may be seen that, as we consider nuclei of increasing charge, the tracks they produce get thicker; and they become more and more 'whiskery' in appearance owing to the increasing number of relatively slow electrons of short range which the fast nucleus knocks out of the atoms it penetrates in its passage through the emulsion. By counting the number of these δ-rays, as we call them, we can identify the charge of the nucleus which produced the track, and thence the kind of chemical element of which it is a representative.

Knowledge of the relative frequencies of the different types of

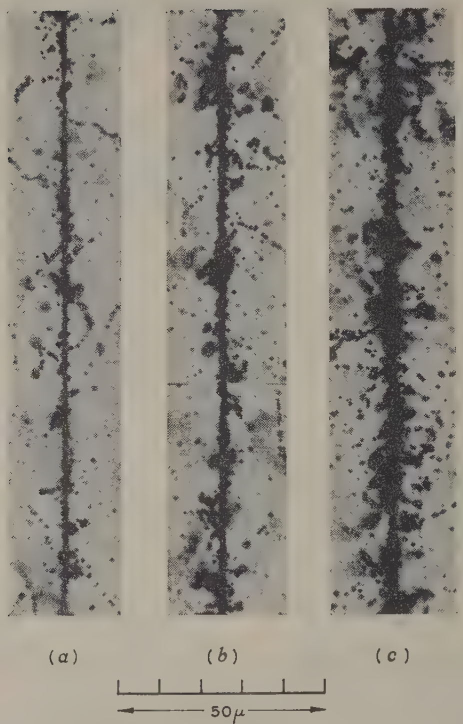


Fig. 1.—Tracks of three heavy nuclei of the primary cosmic radiation moving with velocities approaching that of light.

The different appearance of the tracks with changing nuclear charge is apparent. The spurs are due to electrons ejected from the atoms through which the primary particle passes; by counting the number of spurs in a given length of track, the charge of the primary particle can be estimated.

(*a*) neon (charge 10).
(*b*) silicon (charge 14).
(*c*) calcium (charge 20).

nuclei present in the primary cosmic radiation is very important for our speculations about the origin of the rays, and this involves us in an important technical problem. When a fast nucleus penetrates matter, it sometimes collides with a nucleus of one of the cloud of atoms through which it is passing. If the fast nucleus is a proton, the simplest kind of nucleus we know, it may disintegrate the struck nucleus—knock it to bits—into its component neutrons and protons. On the other hand, if the incoming particle is itself a complex nucleus, it may also be disintegrated into its component neutrons and protons as a result of the collision. Such an event has occurred in the collision in Fig. 2.

As a result of such processes the primary particles of the cosmic radiation, as they enter the earth's atmosphere, begin to make collisions which change their nature. An original nucleus of iron, for example, may fragment and thus be replaced by two or more nuclei of lighter elements of smaller charge. As the primary radiation penetrates into the atmosphere it therefore changes in constitution. The heavier nuclei, being bigger than the lighter ones, are more likely to make collisions. As a result, the charge spectrum is distorted, the proportion of lighter elements being increased and of the heavier diminished. Indeed, few of the heavier elements are able to penetrate the atmosphere to altitudes less than about 90 000 ft because of such collisions. If we want to know the charge spectrum as it was before being modified by such processes, it is clearly necessary to carry our detecting apparatus to such an altitude that the effects of the overlying air are, from this point of view, either eliminated or

Fig. 2.—Fragmentation of an iron nucleus into several parts.

In addition to a number of protons, the impact breaks down the primary nuc into one of oxygen, one of lithium and one of helium. The collision occurs in middle of the left-hand photograph, and the tracks of the three secondary he particles can be separately distinguished in the right-hand photograph, which taken several millimetres from the point of impact, so that the three tracks h diverged sufficiently to be distinguished.

made very small. There are several ways in which this may done.

Rockets are able to reach altitudes of more than 50 mi where they are virtually outside the atmosphere, but they suf from the fact that their time of flight is restricted to a f minutes, and this is too short for the accumulation of suffici data when using the photographic method of detection. T ideal solution would be to carry a large stack of emulsions o satellite, in an equatorial orbit at a height of about 500 mil and to recover the equipment after it had been in orbit about twenty days. Doubtless this will eventually be done, b it must await the solution of the problem of making the recov safely and without damage to the photographic material. In t meantime we have to use balloons (Fig. 3).

It is now possible to make experiments with large balloo of polyethylene or of Terylene, which can carry loads up about half a ton, sometimes more, to altitudes above 100 00 and to maintain altitude for more than 24 hours. In this peri the balloon is likely to drift 1 000 miles in the upper win As an illustration of the kind of problem this imposes, I m

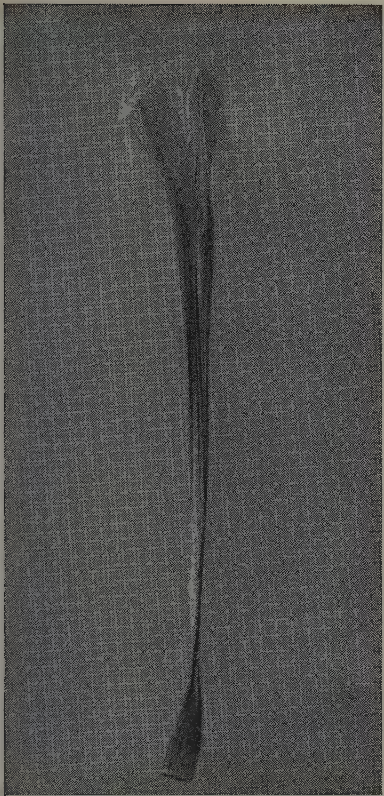


Fig. 3.—A polyethylene balloon at the instant of release.

The balloon is about 200ft long, and only about 0·5% of its volume is filled with hydrogen at ground level.

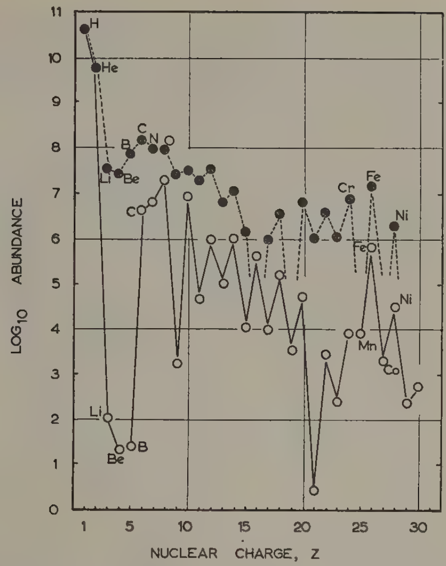


Fig. 4.—Relative abundances of different elements in the primary cosmic radiation and in the general matter of the universe.

Results for cosmic radiation are shown by full circles; for the general matter of the universe, open circles, the two distributions being normalized to hydrogen and helium. The crucial points of difference are the appreciable abundance of lithium, beryllium and boron, and the over-abundance of nickel, iron and other neighbouring elements, in cosmic radiation.

mention that we are at present making plans to launch balloons from the Eastern Mediterranean, possibly from Cyprus, and to recover the equipment, after it has been cut off from the balloon and has fallen by parachute, with naval assistance from somewhere near Malta.

The Charge Spectrum of the Cosmic Radiation

Studies of the constitution of the primary cosmic radiation made by these methods have established the main features of the charge spectrum of the cosmic radiation, and this is illustrated in Fig. 4. The most numerous of the primary particles are protons—hydrogen nuclei. In addition, for every 100 protons there are about 10 helium nuclei—α-particles—and about one heavier nucleus. Among this 1% of heavier nuclei, carbon, nitrogen and oxygen, Nos. 6, 7 and 8 in the periodic table, are prominent; and there are about a third as many of the lighter elements, lithium, beryllium and boron, Nos. 3, 4 and 5. Nuclei of the elements from 10 to 26 are represented with approximately equal frequency, although there is some evidence that the heavier ones, Nos. 20–26, are more numerous than Nos. 14–20. Any nuclei above Nos. 26 or 28, if they occur at all, are very rare. Thus, it has been reported that, in an experiment with counters carried by one of the recent Russian satellites, which was designed to detect the presence of such very heavy nuclei, a single pulse corresponding to the arrival of a single particle heavier than a nickel nucleus was observed throughout the course of the flight.

It may be seen from Fig. 4 that the broad features of the constitution of the cosmic radiation outlined above are very similar to those of the matter of the universe. The relative abundances of different atoms in this general matter of the universe have been inferred from observations on the spectrum of the light emitted from the stars, from studies of the constitution of meteorites, and from terrestrial observations. There also we find a great preponderance of hydrogen and of helium, together with an appreciable proportion of elements up to nickel and iron, but the heavier elements beyond Nos. 26 and 28 are very rare. If this were all the story we might conclude that the particles of the cosmic rays are emitted from a typical star like our sun, of which there are about 10^{11} in our galaxy; it would then remain only to inquire how the particles gained their great speeds. But the situation is more complicated than this, for there are two important ways in which the composition of the cosmic radiation differs from that of average galactic matter.

The first important point of difference is that the composition of the cosmic radiation shows about ten times as much nickel-iron, relative to hydrogen, as galactic matter; nickel-iron is over-abundant in cosmic radiation. Secondly, whereas in the cosmic rays the light elements lithium, beryllium and boron are present in appreciable quantities, they are virtually absent in galactic matter for reasons which we believe we clearly understand. These two features have been of key importance in recent speculations about the origin of cosmic radiation. These speculations owe a great deal to the rapid development, especially in the past five years, of our understanding of the evolution of stars.

According to our present views, a star starts its life as a diffuse accumulation of galactic gas and dust. Under the forces of gravitational attraction between its parts, this matter condenses into a gas ball which becomes hot as a result of the release of gravitational potential energy as the mass contracts. When the temperature of the central regions reaches about 10^6 deg C thermonuclear reactions involving the light elements heavy hydrogen, lithium, beryllium and boron can occur. If such elements are present they are then very rapidly transformed in a

variety of well-known reactions. It is reactions of this kind, involving these elements, which are used in the hydrogen bomb, and which we hope to harness in thermonuclear reactions to provide our main source of power for the indefinite future, before, so we hope, we have consumed all our uranium and thorium in nuclear piles.

It is because of these processes that the light elements lithium, beryllium and boron are present in such low concentrations in the galactic matter. If the cosmic-ray particles originate by ejection from stars, how is it that these light elements appear in considerable concentrations among them? It is suggested that they do not appear directly, but that they are the fragmentation products of heavier nuclei. We have seen an example of an 'event' in which a heavy nucleus, in passing through a photographic emulsion, collides with a nucleus and fragments to give two or three lighter nuclei which may sometimes be lithium or beryllium. Similar processes may occur in the passage of the cosmic-ray particles from their points of origin to the earth, through collisions with the nuclei of the interstellar gas. We can therefore suggest a plausible mechanism to explain the presence of the light elements even though they do not emerge directly from stellar atmospheres where we believe them to be absent. They may be produced as a secondary product of nuclear collisions of heavier nuclei. An important point we shall have to consider later is how much matter the primary cosmic radiation has to traverse to account for the observed intensity of these light elements.

The second suggestive feature of the constitution of the cosmic radiation, to explain which we may again appeal to recent theories of stellar evolution, is the marked abundance of the heavier elements like nickel-iron. We have followed the early stages of stellar evolution which lead to the consumption of lithium, beryllium and boron. Further contraction of the star leads to higher temperatures in the central regions, and reactions can then take place in which the hydrogen is built up into helium. These reactions occur within a central core of the star where the temperatures are of the order of 15×10^6 deg C, and they support the evolution of energy in the star during its main radiating life. It takes about 10^9 years for 1% of the hydrogen to be consumed in the case of a star with the mass of our sun; for more massive stars this period is considerably shorter. When much of the hydrogen in the core has been consumed through the building-up of helium, further gravitational collapse and the corresponding increase in temperature lead to reactions between the helium nuclei, α-particles and the building-up of much heavier nuclei such as oxygen 16 and neon 20. In a small fraction of massive stars this stage is followed by others in which much of the core becomes condensed into nuclei of nickel and iron.

When matter has been condensed into nickel-iron, it is, from the nuclear point of view, in its state of lowest energy, and any further nuclear changes can occur only if energy is provided by other sources. If the central temperature, now of the order of 10^9 deg C, is further increased by gravitational contraction, the nickel-iron nuclei may be dissociated by the electromagnetic radiation, some of which, at these high temperatures, is in the form of γ-rays. This involves the absorption of energy, and a rapid collapse of the core follows.

As a result, the surrounding stellar atmosphere, still containing much hydrogen and helium, now unsupported through the collapse of the core, falls inwards into the region of very high temperatures near the core. At these temperatures the reactions involving the transformation of hydrogen into helium, and others involving the release of neutrons, which at lower temperatures proceed relatively quietly, now take place explosively, in a fraction of a second. A catastrophe occurs in a period of the order of minutes, and a super-nova appears in the sky. Th whole structure of the star is blown to pieces; it flares up brilliance so that its intrinsic luminosity for the first 30 da following the explosion is equal to about 10^{10} suns.

The detailed analysis of these nuclear processes in stell interiors, so briefly outlined, has in recent years been active pursued, and it allows us to account for many important featur in the constitution of galactic matter, including the observe relative abundances of the isotopes of many of the heav elements. These successes give us confidence in the correctne of the essential features of the theoretical speculations. For o present purpose the important point emerges that it suggests source of cosmic radiation which might be expected to inje into the galactic system fast nuclei, much richer in the heav elements like nickel-iron than is the average star such as o sun.

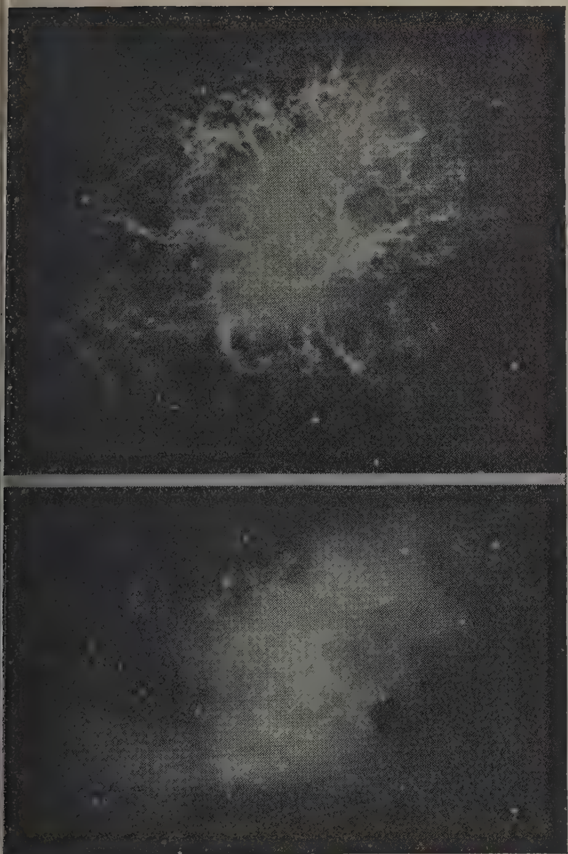
But whilst it appears plausible to assume that material approximately the right kind, apart from the light elemen lithium, beryllium and boron, is being injected into the galact system from super-novae, and that we can thus account for so of the most important features of the constitution of the prima cosmic radiation, the problem arises whether such a mechanis allows us to explain the great energy of the particles, and wheth super-novae occur sufficiently frequently in a galaxy to provi the total energy present in the form of cosmic radiation. A important contribution to these speculations has been made recent studies of the Crab Nebula.

Three super-novae have been observed to occur in our gala in historic times, and studies in radio-astronomy allow us identify about a dozen more. For our present purpose t most important was that recorded in the Chinese annals f A.D. 1054. The Crab Nebula has been identified as the reli of that stellar explosion. This fascinating object is abo 3 000 light-years away, and its present radius is about 5 ligh years. In ordinary light it appears through the telescope as amorphous mass, but when photographed in the light of t hydrogen spectrum (Hα) it has a complicated filamentous stru ture (Fig. 5). The gas is still rapidly expanding at abo 100 km/sec—a speed greater than that of the gas at the cent of an atomic bomb.

The spectrum of the light from some of the central regio of the Crab Nebula shows no line structure like that emitt from excited atoms. It is a continuous spectrum similar to th emitted by accelerated electrons with which we have becor familiar in the past decade, and which we call synchrotron radi tion. Such light is emitted by the beams of particles moving, circular orbits under the action of magnetic fields in the gre synchrotrons. Such accelerated electrons emit light with a fr quency determined by their energy and by the strength of t magnetic field in which they are moving. Further, this emitt light is polarized.

The light from the central regions of the Crab Nebula sho such features. It suggests that there is a marked tendency f magnetic fields to be established parallel to the filamento structures in the expanding gas, and that there are electrons great energy moving in roughly circular orbits round the lin of magnetic force. The observed spectrum shows the presen of electrons of energy $\sim$200 BeV, and it is reasonable to assur some of much greater energy are also present. The presence fast nuclei of even greater energy is even more plausible; th may be accelerated in processes similar to those which give ri to the fast electrons, but they lose their energy much less rapic than do electrons of the same speed, being much less effecti in giving off electromagnetic radiation.

The above considerations allow us to assume that a super-no is able to generate fast-moving atomic nuclei with kinetic energi



(a)

(b)

Fig. 5.—Two photographs of the Crab Nebula taken with the Mount Wilson telescope.

(a) With a red filter transmitting light of Hα.
(b) With a blue filter which passes the synchrotron radiation.
The pattern of the photograph of type (b) changes with the polarization.

in the range which we meet in cosmic radiation, and with a charge spectrum which is richer in heavier elements than the ordinary matter of the universe. It must be remarked, however, that we cannot yet give a satisfactory account of the precise mechanism whereby the particles, either electrons or protons, are accelerated. The tenuous gas in the relics of a super-nova is highly ionized, with associated magnetic fields similar in many respects to the plasmas we employ in present attempts to produce controlled thermonuclear reactions. This is matter in a state with which we are still little familiar, and the theoretical treatment of the interactions of the magnetic fields and the currents, the subject-matter of magneto-hydrodynamics, is very difficult except in highly simplified systems. But there have been optical studies of the Crab Nebula which indicate the presence of rapidly moving light-pulses. The outward movement of the matter in the super-novae, originally due to the force of the explosion, provides a source of energy whatever the precise mechanism whereby it is transmitted to the charged particles.

I have said that the acceleration of protons is easier than that of electrons because they lose little energy in producing synchrotron radiation. The remark needs to be qualified, however, in this sense: It is substantially true at relativistic velocities, for the effective masses of the particles are then effectively defined by their energies; for energies of 200 BeV both electrons and protons have energies about 200 times that of a proton at rest. But at lower velocities a charged particle loses energy by ionization and the rate of loss in passing through a given medium varies approximately inversely as the square of the velocity. This effect makes it more difficult to accelerate protons up to energies of 1 BeV than electrons, if the density of the medium through which the particles are moving is such that ionization losses are appreciable, and any detailed picture of the mechanisms of acceleration will have to account for a satisfactory injection mechanism. But we have good evidence for the presence of electrons of energy 200 BeV, and the presence of fast protons is thus made plausible.

Two features of the primary cosmic radiation now remain to be accounted for: the isotropic nature of the radiation approaching the solar system, and the presence of lithium, beryllium and boron, which are almost certainly absent from a super-nova immediately before it explodes. Why, for example, if indeed super-novae are an important source of cosmic rays, do we not find the particles which reach us tending to come from such a relatively near source as the Crab Nebula? For an approach to this problem we can appeal to some of the recent findings about the structure of the galaxy based on studies of radio-astronomy.

Our galaxy is a spiral nebula which is something like 70 000 light-years across. The stars are most dense near the centre of the system, and the spiral arms are made of 'stars' contained in a disc about 3 000 light-years thick. About 98% of all the matter in the galaxy is contained in the stars, but about 2% of it is in the form of interstellar gas and dust, most of which is confined to a thin flat layer in the plane of the galaxy in a belt about 1 000 light-years thick. This interstellar matter, gas and dust, is very attenuated; it corresponds to a mean density of between about 1 and 10 atoms per cubic centimetre.

But, in addition to these features, studies in radio-astronomy have shown that surrounding the spiral arms of a galaxy there is a halo of gas, getting thinner with distance from the centre of the galaxy. This halo, roughly spherical in form, about the centre of the galaxy, is more than 120 000 light-years in diameter, and the density of matter in it is very much lower than in the disc. Evidence from the halo is derived from studies of the radio emission at a wavelength, for example, of 5 m. The gas and dust in the galactic plane, and in the halo, are not uniformly distributed. As in the Crab Nebula, there appear to be clouds and filaments of matter—irregular concentrations. And, most important, there are irregular magnetic fields associated with these clouds. The average field strength in the galactic disc is estimated to be about 10^{-5} gauss; in the halo it is somewhat less. We attribute the radio emission to synchrotron radiation from electrons. From its observed intensity it is estimated that the number of hydrogen ions and electrons in the halo corresponds to a mean value of the order of 20 per litre.

As a consequence of these magnetic fields in the galaxy and in the halo, a charged particle, even of great energy, cannot pursue a straight course in interstellar space. It will be deflected, and the less energetic the particle the greater will be the deflection in a given field. As a result of the actions of these irregular magnetic fields, charged particles emitted from a super-nova will be subject to random deflections as they move through interstellar space, and their directions of motion will constantly change. Their progress from their source will be similar to that of a gas molecule in the body of a gas—a random walk with sudden erratic changes in direction—the process which leads to diffusion. This allows us to account for the fact that at our point of observation the particles appear to be arriving equally from all directions.

Now, the limited number of super-novae observed in our galaxy and the frequency of their occurrence in neighbouring

galaxies suggest that they occur about once in every 300 years per galaxy. It is reasonable to suppose that they occur most frequently near the centre of the galaxy where the star population is highest, though definite evidence on this point is still lacking. If so, we can assume that most of the cosmic rays arise in the central regions of the galaxy, where the cosmic-ray density will be greatest, and thence that they diffuse outwards towards the periphery of the halo, the density of the radiation becoming progressively less. Eventually some of the particles, especially the most energetic ones, may diffuse out of the galactic halo and enter the regions of space between the galaxies. We do not know the distribution of the intensity of cosmic radiation within the galaxy, and the total amount of energy which it represents, so we cannot yet say whether the frequency of occurrence of super-novae, and the total amount of energy which each injects into the galaxy in the form of cosmic rays, is sufficient to account for the cosmic radiation. But tentative estimates suggest that the magnitudes are of the right order.

Finally, we must return to the question of the lithium, beryllium and boron in the radiation reaching our atmosphere. We have seen that we can account for it as resulting from the fragmentation of heavier nuclei. We can get an idea of how much matter has been penetrated in the following way: Suppose all the fast cosmic rays injected by the explosion of the super-nova were in the form of nickel-iron with no lighter elements at all. Then we find that to reduce the proportion of nickel-iron to the values observed in the primary cosmic radiation would require passage through about 10 g/cm^2 of the interstellar gas—which is largely hydrogen and helium. That is, if all the gas in the regions through which the particles passed on their way from the source to the top of our atmosphere were pressed together it would weigh 10 grammes for every square centimetre. We know that the primary stuff cannot all be nickel-iron; some lighter elements like carbon, nitrogen and oxygen must also be injected, so this is an upper limit.

We can get another estimate by calculating the amount of gas which must have been penetrated to provide the amounts of lithium observed, and this gives a value of about 3 g/cm^2. But we do not know at what stage in the history of the cosmic-ray particles which reach our atmosphere they passed through such an amount of matter. If, for example, they move in the gas of the galactic disc, where the density is relatively high, they will have to penetrate a much smaller distance before making a collision than when moving in the much more tenuous gas in the halo; and a smaller distance still when moving in the expanding gas of the relics of the parent super-nova.

So we cannot yet estimate the time it takes, on the average, for a particle of the cosmic radiation to pass from its point of origin to the earth. If, for example, the diffusion of the cosmic-ray particles were confined to the relatively dense gas in the galactic disc, the traversal of an amount of matter corresponding to the penetration of 3 g/cm^2 of hydrogen would require 10^7 years; if its random motion were confined to the halo, approximately 10^9 years.

So our present state of knowledge leaves many important questions unanswered. Some authors suggest, for example, that there are other sources of cosmic radiation in addition to super-novae. They consider that there is too much hydrogen in the primary radiation to be accounted for if this were the only source. Indeed, we know from recent observations that during some solar flares the sun emits protons of moderate energies, approximately 100 MeV, in large numbers. But it appears more difficult to account for their acceleration to the great energies characteristic of the cosmic radiation than in the case of particles from the super-novae, and the sun appears to emit very few, if any, heavier elements.

Some authors, too, prefer to assume that the diffusion processes take place only in the spiral arms of the galaxy within which the particles are confined by the stronger magnetic fields which exist there. It is an open question also whether the particles of greatest energy of which we have evidence are due to protons. They may be heavier nuclei such as nickel-iron, which would be more easily confined by the general magnetic fields.

So we are at a particularly fascinating stage in the history of the subject. The development of knowledge about the evolution of stars and the generation within them of the chemical elements from the primeval hydrogen, and our increasing knowledge of the structure of the galaxies by studies in radio-astronomy, have greatly increased our insight into the possible mechanisms whereby the cosmic radiation is produced. Certainly much remains to be done, but we believe many of our present speculations must have features not far removed from the truth. In this situation, I cannot do better than end with a remark of Bacon's:

> The universe is not to be narrowed down to the limits of the understanding, as has been man's practice up till now; but rather the understanding must be stretched and enlarged to take in the image of the universe as it is discovered.

621.316.825.2

The Institution of Electrical Engineers
Paper No. 3176 M
Jan. 1960

THERMISTORS, THEIR THEORY, MANUFACTURE AND APPLICATION

By R. W. A. SCARR, B.Sc.(Eng.), Ph.D., Associate Member, and R. A. SETTERINGTON, B.Sc.

(*The paper was first received* 15*th November*, 1958, *in revised form* 18*th July, and in final form* 29*th September*, 1959. *It was published in January*, 1960, *and was read before the* NORTH-WESTERN MEASUREMENT AND CONTROL GROUP 16*th February, and the* MEASUREMENT AND CONTROL SECTION 5*th April*, 1960.)

SUMMARY

Thermistors are components of comparatively recent development, although the phenomenon of conduction in metallic salts was recorded by Faraday over a century ago.

The paper, in a broad survey of the subject, describes the theory of their operation, explains why they, in common with other semiconductor devices, possess a negative temperature coefficient of resistance and develops the expressions which govern their parameters.

Methods of manufacture of the various forms of thermistor are outlined in general terms, and numerous applications of the device are described and discussed. Guidance is given on the approach to circuit problems, and the possibilities of exploiting some of the device's interesting and unusual properties are indicated. The point is made that the thermistor can now be used in many fields as an adequate alternative to components of more specialized application.

LIST OF SYMBOLS

A = A constant.
α = Temperature coefficient.
α_0 = Value of α when $T = T_0$.
B = A constant.
C = Capacitance.
c, D = Constants.
d = A small negative number.
F = A constant.
h = Specific heat, J/g per deg C.
I = Current, amp.
k = Boltzmann's constant.
k'' = A constant (of the order 3×10^{-12}).
K = Dissipation constant.
L = Effective inductance in the equivalent circuit of a thermistor.
μ = Mobility of electrons.
n = Number of electrons in a free state.
N = Number of places occupied by valence electrons.
R = Resistance, ohms.
R_0 = Cold resistance.
R_ϕ = High-frequency resistance of a thermistor.
R_R = Relay resistance.
R_s = Slope resistance.
R_∞ = Value of R when thermistor temperature tends to infinity.
ρ = Density of thermistor material, g/cm^3.
ρ_s = Resistivity, ohm-cm.
t = Time.
T = Absolute temperature.
T_A = Ambient temperature.
T_0 = An arbitrary temperature reference.
T_1 = Thermistor temperature when $V = V_{max}$.
T_2 = Half-temperature.
τ = Thermal time-constant.
θ = A given constant temperature.
v = Volume, cm^3.
V = Voltage.
V_{max} = Maximum value of V (occurs when $dV/dI = 0$).
W_g = Energy gap.

(1) INTRODUCTION

A thermistor is a temperature-sensitive resistor. It has a large negative resistance/temperature coefficient which typically lies in the range 1–5% per deg C. The law of resistance against temperature is usually exponential. Thermistors are made from materials which are classed as semiconductors and in practice these are often a combination of metallic oxides or sulphides.

Conduction in thermistors is purely electronic and, provided there is no change in the temperature of the element, current is proportional to applied voltage.

(2) MATERIALS

Conduction in silver sulphide was first recorded by Faraday in 1837, and the nearly metallic degree of conduction by amorphous carbon is utilized in arc lamps and in brushes for rotating machinery. Triferric tetroxide can also be used as a thermistor material although its temperature coefficient is low compared with that of other materials.

Before 1939, thermistors were made on the Continent from uranium dioxide, from a magnesium oxide-titanium dioxide mixture and from copper oxide. These materials have, in general, been superseded because of inherent difficulties in their manufacture; for example, in order to obtain practical resistivity values from copper oxide it has to be sintered at a temperature of 1 000° C and a pressure of 13 000 lb/in^2 in oxygen.

During the war, developments[1] in the United States using mixtures of manganese, nickel, copper and cobalt oxides started the modern trend in thermistor materials which are sintered in air and are readily controlled in resistivity by varying the proportions of the constituent oxides. More recently, materials in commercial use have been based on mixed ferrites or triferric tetroxide containing traces of lithium oxide. All the materials at present used in thermistors appear to be mixtures of various oxides. In addition to those already mentioned, oxides of vanadium, chromium, titanium and tungsten have been used.

(3) THEORY OF OPERATION

(3.1) Conduction in Oxide Semiconductors

Solid materials may be divided broadly into three groups, conductors, semiconductors and insulators. The division can be made arbitrarily in terms of the resistivities; conductors have resistivities less than 0·1 ohm-cm, semiconductors between 0·1 and 10^9 ohm-cm and insulators greater than 10^9 ohm-cm. The more scientific division of the three groups is made by referring to the value of the energy gap. This is non-existent in conductors, is between 0·1 and 3 eV in a semiconductor and greater than 3 eV in an insulator. It will be appreciated that there can be no rigid division between the various groups.

This is an 'integrating' paper. Members are invited to submit papers in this category, giving the full perspective of the developments leading to the present practice in a particular part of one of the branches of electrical science.

The authors are with Standard Telephones and Cables, Ltd.

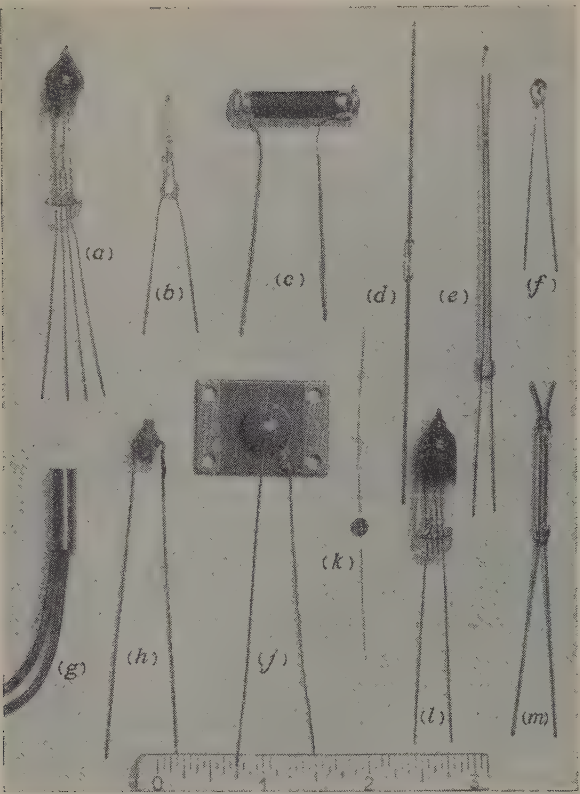


Fig. 1.—Various types of thermistor.

(*a*) Indirectly-heated bead for remotely controlled variable resistor, e.g. in automatic gain control.
(*b*) Directly-heated bead for time delay, a.g.c. circuits or current surge suppression.
(*c*) Rod-type for current surge suppression.
(*d*) Directly-heated bead for microwave power measurement.
(*e*) Directly-heated bead, probe-type thermometer.
(*f*) Directly-heated bead with a wide range of applications.
(*g*) Bead (*f*) mounted in metal case.
(*h*) Disc-type for temperature measurement, control and compensation (up to 120° C).
(*j*) Disc type, as for (*h*), mounted for surface measurements.
(*k*) Directly-heated bead, skin thermometer.
(*l*) Directly-heated bead, a more power-sensitive version of (*b*).
(*m*) Directly-heated bead for anemometry and thermal conductivity measurements, e.g. vapour-phase chromatography.

The oxide semiconductors form a class of materials in which conduction is electronic as opposed to ionic and which, as far as their temperature sensitivity is concerned, have a good deal in common with intrinsic monocrystalline semiconductors such as germanium and silicon. The structure and the mechanism of conduction in a monocrystalline semiconductor is reasonably well understood but this is far from being the case with an oxide semiconductor. It is useful to enumerate the ways in which the physical properties of an oxide semiconductor differ from those of a monocrystal:

(*a*) The oxide semiconductors are made up of granules, each of which is a monocrystal of a definite structure, but the very large number of grain boundaries must also play an important part in the semiconductor action.

(*b*) The bonds between atoms are partly ionic and partly covalent rather than completely covalent.

(*c*) Conduction depends on physical imperfections rather than on chemical impurities. While impurities can modify the behaviour of an oxide semiconductor and a high order of chemical purity is necessary in its constituents, such requirements are not nearly as stringent as in monocrystalline materials.

(*d*) There are two forms of conduction and these are recognized as resulting from excess of electrons or deficiency of electrons, i.e. holes. These correspond, not to acceptor and donor impurities, but to two types of lattice defects, Frenkel defects and Schottky defects. In a Frenkel defect there is an extra ionized metal atom in the lattice which is not bound to an oxygen atom and the electrons associated with this atom can take part in conduction. The result is an excess or *n*-type semiconductor. In a Schottky defect there is a positively ionized divalent metal atom missing from the lattice and to maintain electrical neutrality two electrons must also have been removed. This results in a deficit or *p*-type semiconductor. The *n*-type semiconductor corresponds to a deficit of oxygen and the *p*-type to an excess of oxygen. This explains the importance of the oxygen content of the atmosphere in which thermistors are made and the importance of the heat treatment which produces the defects.

This qualitative summary is based on an excellent survey of the subject given by Goudet and Meuleau.[2] No quantitative theory appears to exist which can adequately explain the behaviour of the oxide semiconductors.

(3.2) The Resistance-Temperature Equations

The relation between the number of charge carriers and the temperature can be obtained by an argument on the following lines.[3] It is assumed that in the crystal lattice there are N places occupied by valence electrons and that at an absolute temperature, T, there will be n of these N electrons in a free state. An activating energy, W_g, is required to excite the electrons. This corresponds to an effective energy gap and has a value of about 0·5 eV for a typical thermistor. (A similar argument could be used for holes.) In thermodynamic equilibrium

$$n^2/(N-n) = F\exp(-W_g/kT) \quad . \quad . \quad . \quad (1)$$

where F is a constant or a function which is slightly dependent on temperature. If n is much smaller than N,

$$n = \sqrt{(FN)}\exp(-W_g/2kT) \quad . \quad . \quad . \quad . \quad (2)$$

The conductivity is equal to the product of n and the mobility μ, of the electrons. Mobility will probably vary with temperature according to a power law which will depend on the materials concerned:

$$\mu \propto T^d \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (3)$$

where d is a small negative number.

From the above, the equation quoted by Becker, Green and Pearson[1] for the resistivity of thermistor materials is apparent

$$\rho_s = AT^{-c}\exp\frac{D}{T} \quad . \quad . \quad . \quad . \quad . \quad (4)$$

where A, c and D are constants. In practice, D usually lie between 1 500° K and 6 000° K, c is a small positive or negativ number or may be zero and A can have a very wide range o values.

The above is usually expressed as a resistance equation an reduced to the form

$$R = R_\infty\exp\frac{B}{T} \quad . \quad . \quad . \quad . \quad . \quad . \quad (5)$$

where B is a constant which is typically of a value similar to D. This is accurate enough for a number of practical application when temperature changes are small compared with the absolut temperature. An empirical form of the thermistor resistance temperature equation has been given by Bosson *et al.*[4] and this :

$$R = R_\infty\exp\frac{B}{T+\theta} \quad . \quad . \quad . \quad . \quad . \quad (6)$$

where θ is a constant temperature. This equation is shown t fit the measured characteristics of two manufacturers' thermisto with fair accuracy over a range of temperature.

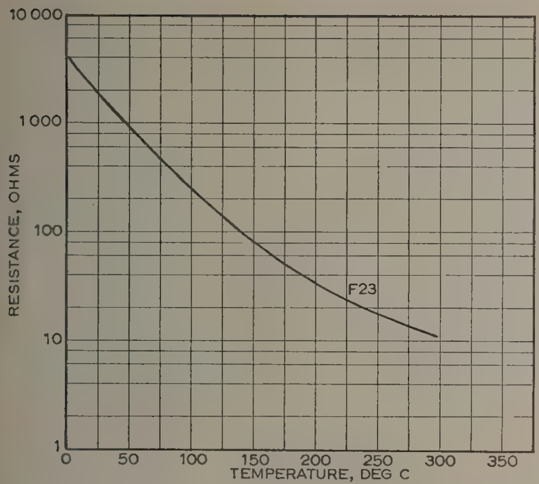


Fig. 2.—Resistance/temperature characteristic of a bead thermistor.

A typical resistance/temperature characteristic for a bead thermistor is given in Fig. 2.

(3.3) Dissipation Constant

The incremental dissipation constant—the power required for a unit temperature rise—is a parameter of importance in many thermistor problems. The dissipation constant is dependent on the nature of the environment of the thermistor and on how the thermistor is mounted; the conditions under which it is measured must therefore be specified. Strictly, the dissipation constant is not constant but depends to a small degree on the temperature at which it is measured.[1]

When the temperature difference is large, the power dissipated is not a linear function of temperature difference and Smith[5] has given a formula for power loss:

$$P = K[\Delta T + k''(\Delta T)^6] \quad . \quad . \quad . \quad . \quad (7)$$

where k'' is of the order 3×10^{-12}. In practice the dissipation constant, K, may range from $10\,\mu$W per deg C for a bead *in vacuo* to 50 mW per deg C for large rods or for discs mounted on a heat sink.

(3.4) Time-Constant

The time taken for the excess temperature of the thermistor to drop on cooling to $1/\varepsilon$ of its initial value is the thermal time-constant, τ, and to calculate τ one must have a knowledge of the thermal capacitance (i.e. mass and specific heat) of the thermistor as well as its dissipation constant. The density of thermistor material is about $4\cdot5\,\mathrm{g/cm^3}$ and its specific heat about $0\cdot7$ J/g per deg C. The thermal time-constant is given by

$$\tau = \rho vh/K \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (8)$$

Table 1 shows values of τ and K for various types of thermistor. Fig. 3 gives a cooling curve for a rod thermistor; in this instance the cooling law is exponential, but with other forms of construction, for example when the thermistor is surrounded by a solid material such as glass or resin, marked departures from a simple exponential law can occur.

Table 1

THERMISTOR CONSTANTS

Thermistor	Construction	K	τ
		W per deg C $\times 10^{-3}$	sec
A	Bead in gas-filled bulb	0·28	1·5
CZ1	Rod	18	140
K	Disc	15	90
KB	Disc on mounting-plate attached to heat sink	100	33
G	Bead in glass	2	24

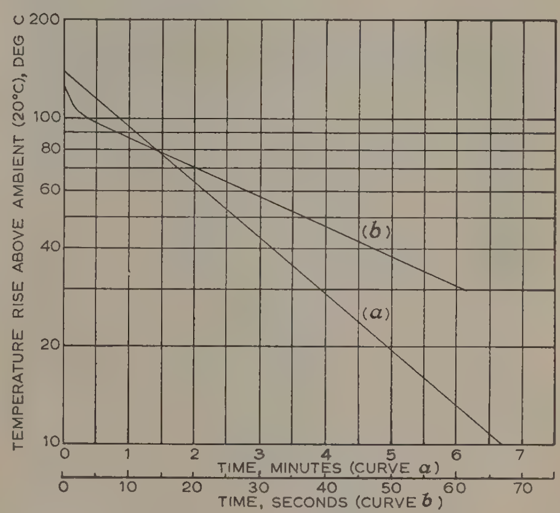


Fig. 3.—Cooling curves.

(*a*) Rod thermistor.
(*b*) Unit encapsulated in resin.

(3.5) Temperature Coefficient and Half-Temperature

The temperature coefficient, α, is defined as

$$\alpha = \frac{1}{R}\frac{dR}{dT} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (9)$$

and if the simplest form of the thermistor resistance-temperature relationship, eqn. (5), be assumed,

$$\alpha = -B/T^2 \quad . \quad . \quad . \quad . \quad . \quad . \quad (10)$$

It is sometimes useful to know the temperature change, ΔT_2, which will halve or double the thermistor's resistance compared with its value at an arbitrary reference temperature, T_0:

$$\Delta T_2 = \frac{\log_\epsilon 2}{-\alpha_0 \pm \dfrac{\log_\epsilon 2}{T_0}} \quad . \quad . \quad . \quad . \quad . \quad (11)$$

where α_0 is the value of α when $T = T_0$; the plus sign applies to doubling the resistance and the minus sign to halving it.

(3.6) The Voltage Maximum

The static voltage/current characteristics, Fig. 4, have a voltage maximum, V_{max}, and it may be shown that this is dependent only on the dissipation constant, the cold resistance, the ambient temperature, and B:

$$V_{max} = \sqrt{[R_0K(T_1 - T_A)]}\exp\frac{B(T_A - T_1)}{2T_1T_A} \quad . \quad (12)$$

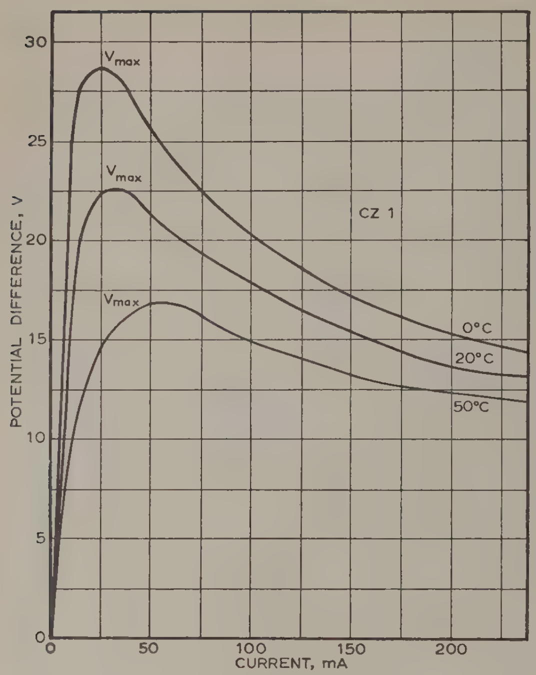


Fig. 4.—Voltage/current characteristics of a rod thermistor at various ambient temperatures.

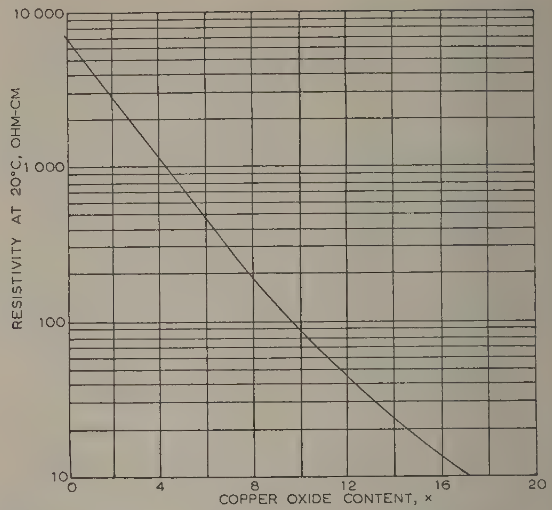


Fig. 5.—Variation of resistivity at 20° C with copper-oxide content.
$Mn_2O_3 : NiO : CuO$
$80 : 20 : x$

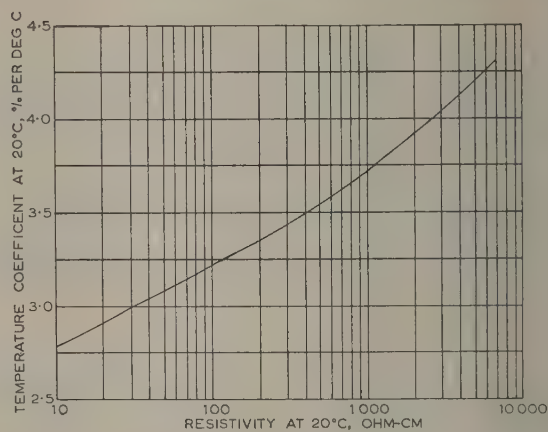


Fig. 6.—Temperature coefficient versus resistivity using material shown in Fig. 5.

where $$T_1 = \frac{B}{2}\left[1 - \sqrt{\left(1 - \frac{4T_A}{B}\right)}\right] \quad . \quad . \quad . \quad (13)$$

which is the temperature of the thermistor when $V = V_{max}$.

Only materials with a negative temperature coefficient of resistance will show a voltage maximum.

(4) THERMISTOR MANUFACTURE

The normal types of thermistors may conveniently be grouped into four basic types, rod, disc, flake and bead, according to the method used to obtain the final shape of the semiconductor body (Fig. 1). In each case the basic materials, usually metallic oxides, are intimately mixed and ground to a homogeneous fine powder in a ball mill or similar device. This powder is then bound by a suitable organic binder, sintered by firing after being shaped to the required form, and dried. The unit is then aged to achieve electrical stability. The ageing process is lengthy and precise in the case of thermistors designed for exact thermal measurement applications. For example, a thermistor may be measured at 20° C, stored for a week at 100° C and rejected if it shows a resistance change of more than 1% during this period.

The resistivity of the material is determined mainly by the proportions of the ingredients and it is usual to maintain a careful control of the materials by chemical analysis of the mixtures. Minor variations in the resistivities of the final product are corrected by slight variations in the firing temperature, which varies between 1 000 and 1 350° C, although the permissible range for any one material may be as little as 10° C. The nature of the atmosphere around the thermistor bodies during the sintering process varies with the starting materials, and it may also be necessary to control the rates of heating and cooling of the bodies, as well as the maximum temperature.

Fig. 5 shows how the resistivity of the semiconductor material decreases with increasing copper-oxide content, and Fig. 6 shows how the temperature coefficient is related to the resistivity.

Rod-type thermistors are usually made by combining the powdered oxides with an organic binder to form a plastic body, extruding the material through a die, usually of circular cross section, and cutting to length. The bodies are dried and fired slowly to avoid cracking and distortion. The processes used are analogous to those used in the ceramic industry. The sintered rods have metallic contacts applied to the ends, either by metal spraying or by sintering on a metallic paste, and the units are completed by pressing on wired end-caps or by soldering on lead-wires. The solder used for this process is usually of a higher melting-point than the normal electrical quality in order to be able to use the full temperature range of the material.

Disc- and washer-shaped bodies are made in machines similar to those used for making medicinal tablets. The dry powder is usually bound and, when dry, recrumbled to grains which flow more easily through the machine than would a fine amorphous powder. The resultant coarse powder is fed automatically into

die where it is formed at high pressure into a relatively hard disc and then automatically ejected. Metallic contacts are applied, as for rod thermistors, and the units may have lead-wires or base-plates attached. With this type of thermistor it is possible to obtain an economical output of close-tolerance units as the resistance can be adjusted by removing small areas of metal from one of the faces of the disc.

Flake-type thermistors, used as sensitive infra-red detectors, are made by applying to a glass plate a coating of thermistor material suspended in an organic binding solution and, after drying, cutting the film into flakes before firing. This type is not currently in production in Europe.

Thermistor beads are components of a wide range of devices, the size of the bead and the diameter of its lead-wires depending on the final unit. The basic method of manufacture is the same in each case. The choice of materials for the lead-wires is limited to materials which

(*a*) Have a higher melting-point than the sintering temperature of the thermistor.
(*b*) Have a coefficient of thermal expansion similar to that of the thermistor.
(*c*) Are resistant to oxidation in the case of thermistors sintered in an oxidizing atmosphere.

These considerations usually limit the choice to platinum or one of its alloys. A jig is used to hold the lead-wires (0·001–0·004 in diameter) parallel under slight tension and spaced 0·003–0·015 in apart. A slurry is made of the powdered oxides by mixing them with a dilute organic binder and is applied manually to the wire forming strings of beads. The diameter is carefully controlled by inspection on a projection magnifier. The diameter varies between 0·007 and 0·03 in depending on the type of thermistor required. After drying, the strings of beads are sintered and the lead-wires are cut to produce the individual beads. There are three possible methods of cutting the leads (Fig. 7) giving alternative mountings in the complete

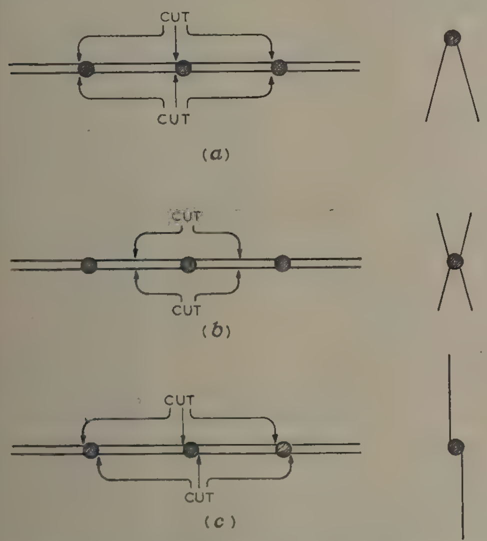


Fig. 7.—Diagrammatic representation of how beads are manufactured and cut for various types of assembly.

thermistor. The individual beads are pre-tested under conditions of controlled temperature and are stored until needed for final assembly. In order to produce a stable thermistor, it is essential to protect the bead from moisture; this is done by sealing it in solid glass or in a glass bulb. The bulb is either evacuated or filled with a carefully dried gas.

Bead-type thermistors may be directly or indirectly heated. Directly-heated thermistors use beads cut as in Fig. 7(*a*). The lead-wires are welded to heavier wires of a glass-sealing material, such as Cunife, and pinch-sealed into a bulb. A 0·007 in bead in an evacuated bulb gives a power sensitivity of 0·02 mW per deg C, whilst a 0·03 in bead in a gas-filled bulb gives a power sensitivity of 0·5 mW per deg C. Indirectly-heated thermistors use beads cut as in Fig. 7(*b*). The bead is enclosed in a small heater made from a wire having a low temperature coefficient of resistance, from which it is insulated by a special cement having high electrical and low thermal resistance. This sub-assembly is welded to a four-wire stem, the double connection of the thermistor lead-wires lending rigidity, and is sealed into a glass bulb (Fig. 8). Power sensitivities vary between 0·05 mW

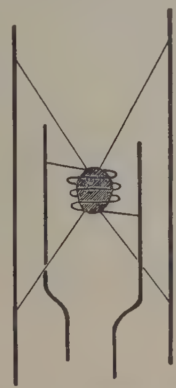


Fig. 8.—Construction of indirectly-heated thermistor.

per deg C for a small assembly in an evacuated bulb and 2 mW per deg C for a large assembly in a gas-filled bulb.

The methods of manufacture of the many special-purpose thermistors, such as manometers, anemometers, r.f. power-measuring thermistors and thermistor thermometers, are basically the same as those described above. Skin thermometers, for example, are made from beads cut as in Fig. 7(*c*), coated with glass and sealed to a copper disc; while for probe-type thermometer the bead, cut as in Fig. 7(*a*), is sealed in glass at the tip of an elongated bulb.

(5) APPLICATION OF THERMISTORS TO THERMAL MEASUREMENT, CONTROL AND COMPENSATION

In almost every field of scientific endeavour and industrial development the thermistor plays an important part as a temperature-sensing element, in the measurement of the thermal conductivity of a medium and in the measurement of heat radiation.

(5.1) Temperature Measurement and Control

The thermistor has the following advantages for temperature measurement:

(*a*) A large temperature coefficient.
(*b*) Small size.
(*c*) An ability to withstand electrical and mechanical stresses.
(*d*) A wide operating temperature range from well below[6] 0° C to 200° C or higher for special types.
(*e*) A wide range of resistance values.

Disadvantages are a non-linear resistance temperature characteristic, the need for auxiliary apparatus to obtain a temperature reading and a reputed inability to maintain its calibration.

The temperature coefficient (eqn. 10) is very dependent on temperature. Examples of its variation for a typical thermistor are given in Table 2.

Table 2

VARIATION OF TEMPERATURE COEFFICIENT WITH TEMPERATURE

Absolute temperature	Resistance	Temperature coefficient
deg K	kΩ	%
200	150	6·7
273	4·1	3·6
300	1·7	3·0
373	0·25	1·9
473	0·034	1·2

By using resistors in series and parallel with the thermistor[7–10] the characteristics can be adjusted to give a linear indication of temperature to a degree of accuracy which will depend on the temperature range to be covered and the temperature coefficient required. For example, with $\alpha = -4\%$ per deg C and a network with a temperature coefficient of $-0{\cdot}08\%$ per deg C, the departure from linearity would be about 0·02° C in 10° C. In general, the characteristic can be shaped by a series and shunt resistor to approximate to any desired curve. To most curves of a simple geometric form, a perfect fit can only be made at a maximum of three points, unless more complex networks using more than one thermistor are allowable.

The accuracy with which temperature may be recorded will depend on:

(*a*) The measuring circuit.
(*b*) Whether absolute temperature or a small temperature difference is required.
(*c*) Whether short- or long-term stability is needed.
(*d*) The construction of the thermistor, e.g. bead or rod.
(*e*) The manufacturing process and the treatment (e.g. ageing) of the thermistor after manufacture.

To record temperature accurately with a thermistor, a Wheatstone bridge with the thermistor as one arm is preferable, but for imprecise measurement simpler arrangements (e.g. a battery, meter and thermistor in series) may be used. It is possible to incorporate the shaping network in the arms of the bridge[10] so that the bridge gives a linear indication of temperature over a limited temperature range. For the measurement of temperature difference, as opposed to absolute temperature, two thermistors are required and these can form two arms of the bridge, but their temperature coefficients must be matched over the operating temperature range.

The literature[1, 11–19] contains a number of references to results which have been obtained in practice for thermistor stability, but it is difficult to draw any very definite conclusions from these as they are to some extent conflicting. The following conclusions can, however, be drawn:

(*a*) Measurement of a small temperature difference over a short period of time can be made to a high order of accuracy; 0·0002° C has been claimed[14] in a temperature difference of 0·01° C. A short-term resistance stability of 5 parts in 10^6 (corresponding to a temperature accuracy of about 0·0001° C) has been observed.[13]

(*b*) For larger temperature differences over a short period of time Beck[12] quotes a figure of the order of 0·01° C in a temperature difference of 6–12° C.

(*c*) Long-term stability will depend on the previous history or ageing of the thermistor and may be affected by thermal shock or large temperature changes. The thermistor should be cycled over a temperature range in excess of its working range prior to calibration. No conclusions about thermistors in general can be drawn regarding long-term stability as the factors affecting performance are too many and too varied.

Fig. 9 shows the result of ageing a batch of 50 thermistor over a period of one year. From these results a long-term stability of ±0·2° C seems to be a reasonable expectation. Th stability could in fact be nearer ±0·1° C if there were calibratio errors which cancelled the apparent thermistor drift.

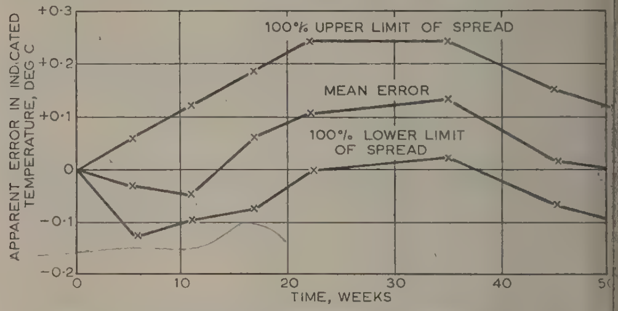


Fig. 9.—Ageing characteristics of 50 thermistors type M52.

Thermistors cycled once to 100° C between each test and stored at room temperature measured at 20° C ± 0·1° C.

Alternative methods of temperature measurement which ar commonly used are glass thermometers, platinum resistance wire thermometers and thermocouples. Each instrument ha advantages and, in general, a thermistor is chosen where remot indication is required, working conditions would result in glass thermometer being broken, small size is essential, or wher a small temperature difference is required. A platinum-resis tance thermometer is chosen where long-term stability is of th highest importance or where the temperature is too high for thermistor. A thermocouple is an alternative to a thermisto in many cases, but its disadvantages are that it measures onl temperature difference and that the associated electrical equip ment may need to be more sensitive. Because the mercury (o alcohol) in a glass thermometer is direct reading, it is usually th first choice for temperature measurement and the alternatives ar employed only for the reasons given above or for other specia reasons.

The delay in a thermistor's response to a sudden temperatur change will depend on its construction and the nature of it environment[20] and may vary from a fraction of a second t several minutes. Speed of response is one of the factors whic determines whether a bead or disc thermistor is most suitable fo a specific job. A bead thermistor will respond more quickl but is less robust and is more subject to self-heating from th current in the measuring circuit. Rod thermistors are no generally recommended for temperature measurements.

The thermistor is a versatile indicator of temperature, as ma be seen from the wealth of its applications; these range fror the measurement of soil,[18] water[19] and air[21, 22] temperature through the measurement of bearing temperature in larg machines[23] to the inclusion of a thermistor inside the bor of a hypodermic needle to measure subcutaneous bloo temperature.[24]

Most temperature-control applications differ from tempera ture-measurement applications only in the way in which th output or error signal from the thermistor is handled. In contro applications the error signal from a thermistor is amplified an either operates a relay[32, 33, 34] on an on-off basis, or contro the conduction period of a thyratron, hard valve or transisto as a proportionate control.[35] By simple analogue compute techniques the rate of change of temperature as well as th absolute error may be taken into account, and fluctuatior about the mean value reduced. Such a technique is used i

the control of block temperature in pressure die-casting machinery.[36]

A temperature alarm system can be made to be much simpler than an accurate control system. Reference to Fig. 4 shows that V_{max} decreases with temperature, and if a constant voltage is applied to a thermistor there will be a sudden increase in current at a critical temperature which is sufficient to operate a relay on, for example, a fire alarm or a standby warning on a radio receiver.[32]

(5.2) Temperature Compensation

The thermistor has the advantage of a relatively large resistance/temperature coefficient so that its characteristic may be shaped to be the same or the opposite of that of the circuit element to be compensated. Another advantage of a large temperature coefficient is that it tends to reduce the loss of the compensating network. Considerable ingenuity may be needed in using a thermistor where a device with a positive temperature coefficient would be more suitable, and this makes the tungsten filament lamp a serious rival to the thermistor in some applications.

Temperature compensation of wire-wound devices such as meter movements[25, 26] and cathode-ray-tube focus and deflection coils are typical examples of the positive temperature coefficient of copper being balanced by the negative coefficient of a thermistor-resistor network. The temperature dependence of the frequency of an oscillator, whether LC[27] or crystal[28] controlled, may be reduced by the use of a thermistor.

Most elaborate compensating systems[29, 30] (Fig. 10) are needed for long cables where the attenuation/frequency characteristic is a function of temperature, and several thermistors may be used in the compensating filters of a repeater amplifier. A pilot channel provides a special signal which is compared with a reference voltage. The resultant error voltage is applied as bias to a thermistor. The resistance of the thermistor affects the gain and acts to adjust it to the desired value. The use of a feedback system of this sort in a line containing a number of repeaters will result in the build-up of low-frequency oscillations unless the time-constants of the thermistors are long compared with other time-constants in the system. Also, the attenuation/phase characteristic of the thermistors at low frequencies may have to correspond to those of a simple RC network. Special directly-heated thermistors with a high ratio of time-constant to mass have been developed in order to obtain the required stability and sensitivity.

Semiconductors in general are noted for their large temperature coefficients and are often compensated by a thermistor, for example in transistor circuits[31] and in photographic exposure meters using selenium cells.

(5.3) Measurement of Thermal Conductivity

If a thermistor is heated directly by an electric current, its resistance will depend upon its heat loss to the surroundings and therefore on the composition and velocity of the medium in which it is situated. Thus, thermistors can be used to determine the composition of a gas in gas chromatography,[37, 38, 39] for the measurement of pressures down to 10^{-6} mm Hg[40] or the control of a vacuum furnace.[41] They may indicate rates of flow, the level of a liquid and wind speed.[21, 42, 43] The thermistor's resistance will also be affected by ambient temperature and some form of compensation or correction is usually necessary, though, if the thermistor is run hot, at say 150° C or more, correction for small ambient temperature changes is not always essential.

The bead thermistor has several advantages over a hot-wire instrument: it is smaller, requires less power, is more temperature sensitive and may be less subject to corrosion.

(5.4) Measurement of Radiation

Thermistors can measure infra-red radiation, but have to some extent been superseded by recently-developed semiconductor compounds such as indium antimonide or lead telluride. The mechanism is again one of a temperature rise causing a resistance change, whereas the semiconductor compounds depend on the generation of minority carriers having an appreciable lifetime. Nevertheless, specially mounted bead-type thermistors are suitable for simple portable instruments, for example in the measurement of solar radiation,[44] and flake thermistors[45] have been used in bolometers.

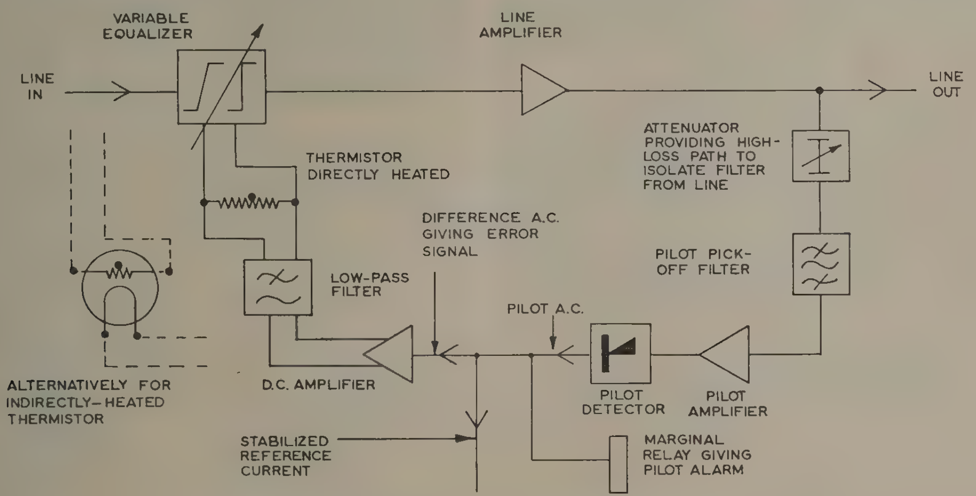


Fig. 10.—Thermistor in line repeater amplifier (pilot regulator).

(6) APPLICATION OF THERMISTORS TO THE MEASUREMENT AND CONTROL OF ELECTRICAL QUANTITIES

The thermistor is very well established as a power-measuring device; that it can be used in the measurement of voltage, current

and impedance in several novel ways is perhaps less well known. It can also play a vital part in a.g.c. systems, volume limiters and expanders, and voltage stabilizers. It can act as a surge suppressor and provide a delay.

All these applications depend on a change in resistance caused either by the signal to be measured or by the control signal or by the signal to be controlled, so that the power sensitivity of the thermistor will be important.

In some voltage-stabilizer circuits, for surge suppression applications and in delay circuits the operation depends on the thermistor having a negative temperature coefficient. A positive-temperature-coefficient device would be satisfactory in a number of these applications, and the fact that thermistors are used is mainly because they have, or can have, a high temperature coefficient, a high power sensitivity and are robust and compact.

(6.1) Measurement of Electrical Quantities at Audio and Radio Frequencies

The indirectly-heated thermistor can act as an a.c.–d.c. transfer device:[46, 47] a known proportion of the alternating voltage or current to be measured is applied to the heater winding of the thermistor and the resulting change in bead resistance is compared with the corresponding result when d.c. is applied to the heater. To compensate for changes in ambient temperature a bridge circuit with thermistors in two arms of the bridge is commonly employed. Limitations on accuracy are imposed by the degree of matching obtainable between the thermistors, at low frequencies by the thermal time-constant of the bead, and at high frequencies by the inductance of the heater winding. A radio-frequency milliameter has been described[48] which will work up to a frequency of 10 Mc/s with a negative error of 4%. Basically these are all power measurements.

For the measurement of a low conductance at a high frequency a bridged-T network is sometimes used. In such cases the resistor which bridges the T must be variable and stray capacitance and inductance effects impose limitations on the component. The r.f. resistance of a thermistor which bridges the T is controlled by the direct current passing through it, and this bias is automatically adjusted to obtain the required balance.[49] Accurate results up to a frequency-resistance product of 50 Mc/s × kilohms are claimed, the upper limit being determined by lead inductance and the shunt capacitance of the bead. This is an example where the thermistor is used as a control device rather than as an indicator of power.

In general, the high-frequency limitations of a thermistor will depend on the application and will be influenced by the following factors:

(*a*) Heater inductance in the case of the indirectly-heated thermistor.
(*b*) For directly-heated thermistors,[50] shunt capacitance and/or lead inductance, but in some cases these effects may be cancelled and distributed capacitance will determine the upper frequency limit.

The effect of distributed capacitance is to reduce the real component of the bead impedance in a way which may vary from bead to bead and which is influenced by the method of manufacture. By the use of suitable manufacturing techniques thermistors giving satisfactory operation at several hundred megacycles per second are practicable. This is one facet of the thermistor's behaviour which could profitably be investigated in greater detail.

A distinction is drawn between the measurements described in this Section, where the thermistor's heater or bead impedance must be known accurately, and those in the next Section, where it need not be known.

(6.2) Power Measurement in Waveguides and Coaxial Systems

Thermistors are commonly used for the measurement of power at radio and microwave frequencies.[51–54] It is the usual practice to cancel the reactive component of the thermistor impedance by a tuning adjustment and the real component does not have to be equal to the low-frequency value; all that is required is for the thermistor to be a good match and matching is adjusted by a d.c. bias on the thermistor (usually a bead type).

The design of the thermistor mount is, however, a factor of major importance in microwave equipment. Reasonably accurate results can be obtained down to a wavelength of 3 cm, but below this a significant proportion of the incident power is dissipated in the leads to the thermistor bead[55] owing to skin effect and this can lead to errors of up to 50% at millimetric wavelengths. It would appear that, if a thermistor is to be used at Q-band frequencies, a new form of construction will have to be found.

For the measurement of power, the thermistor forms part of a bridge circuit and it is convenient to make this bridge give a direct indication of power. Several methods of achieving this are given in the literature.[10, 54, 56]

(6.3) Control of Electrical Quantities, Including Use in A.G.C. Systems, Volume Limiters and Expanders

Many automatic-gain-control systems have thermistors as an element in the feedback path to maintain a constant output level. Output stabilization is achieved because any tendency of the output level to increase results in a decrease in thermistor resistance and hence in an increase in negative feedback. The thermistor may be incorporated in a bridge or form part of a T- or π-network (see Fig. 11). Fixed-frequency Meacham bridge[57]

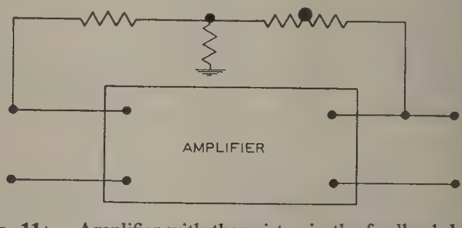


Fig. 11A.—Amplifier with thermistor in the feedback loop.

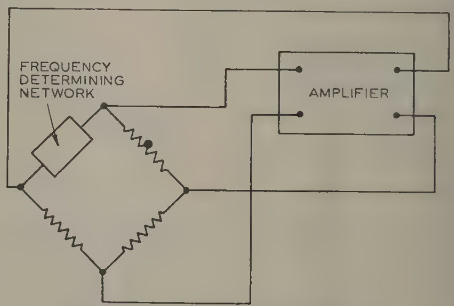


Fig. 11B.—Meacham bridge oscillator.

oscillators are well known and a number of variable-frequency amplitude-stabilized oscillators have been described.[58–61] The application of this principle to the temperature compensation of the characteristics of a telephone line has been mentioned in Section 5.2.

It is not essential for the thermistor to be included in the feed-

back loop, and systems have been described where it forms part of the amplifier[62] or line circuit;[63] such arrangements can be classed as volume limiters or expanders.

Precision alternating voltage stabilizers[64] can have a thermistor as a sensing element by utilizing the fact that, if the alternating voltage to be stabilized is applied to a bridge containing a thermistor, the bridge is balanced at only one voltage and the error signal changes sign at that voltage (see Fig. 12). At a

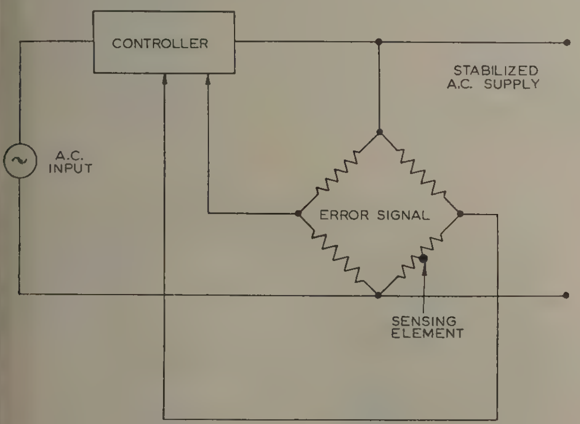


Fig. 12.—Voltage stabilizer.

frequency of 50 c/s, the thermistor temperature is liable to follow to some extent the periodic variations in voltage, and this results in a 50 c/s signal with some phase shift and the production of harmonics of 50 c/s. These effects can limit the precision of the stabilizer; their magnitude can be determined from an analysis on the lines given by Ekelof *et al.*[65]

A thermistor, either indirectly heated or controlled by a current outside the frequency range of the signal band, can form part of a remotely-controlled noise-free variable resistor or attenuator. A programme-fading circuit working on this principle has been described by Whitehead,[66] and a bridge-T circuit is mentioned in Section 6.1.

(6.4) Delay and Surge Suppression

Thermistors behave in a circuit as if they have a large inductive reactance; this is equivalent in physical terms to saying that they have a large thermal mass combined with a negative temperature coefficient. This property of the thermistor enables one to delay the operation of a current-operated device such as a relay[67] or to suppress a switching surge, for example in circuits employing lamps or valve heaters,[68] in starting small electric motors[69] and in suppressing telephone bell tinkle.[70] Probably more thermistors (or equivalent devices known under various trade names) are used in television and radio receivers employing series-connected valve heaters than in any other field.

There is an upper limit to the magnitude of the surge which can be suppressed using one homogeneous thermistor element and this is determined by a maximum mass, above which the thermistor is liable to shatter due to transient thermal stresses. This limit at present corresponds to a peak instantaneous power dissipation of the order of 500 watts. Operation of thermistors in parallel is dangerous because imperfect matching can result in one of the thermistors taking most of the load.

Relay delay circuits using thermistors are common, but a high order of accuracy under conditions of varying ambient temperature should not be expected. By providing an extra change-over contact on the relay which removes a shunt across the energizing

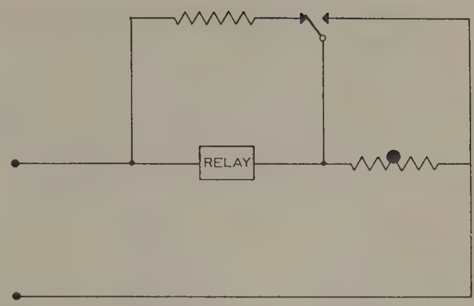


Fig. 13.—Relay delay circuit with an anti-chatter contact.

coil and then short-circuits the thermistor (see Fig. 13), relay chatter can be prevented and the thermistor will have more time to cool before the next cycle.[67] Delays from a fraction of a second to over a minute are possible. Circuits of the ring-counter type may be built for sequential switching.

(7) THE THERMISTOR AS A CIRCUIT-ELEMENT

In analysing steady-state and transient problems in circuits containing thermistors, it is useful to simulate the thermistor's behaviour by an equivalent electrical circuit. Providing the thermistor's temperature remains constant, it behaves as a simple resistance. For small temperature changes (a few degrees Celsius) about a mean value, it can be represented by a comparatively simple 3-element network. For large temperature changes, a limited number of problems can be solved but an equivalent circuit is not practicable.

(7.1) Small-Signal Equivalent Circuit

A number of authors[5, 71, 72, 74, 78, 79, 80] have deduced equivalent circuits for the thermistor, and one containing two resistive elements and an inductive element, Fig. 14(*a*) seems to be preferred.

The value of R_s, Fig. 14(*b*), is determined by the slope of the static voltage/current characteristic at the operating point

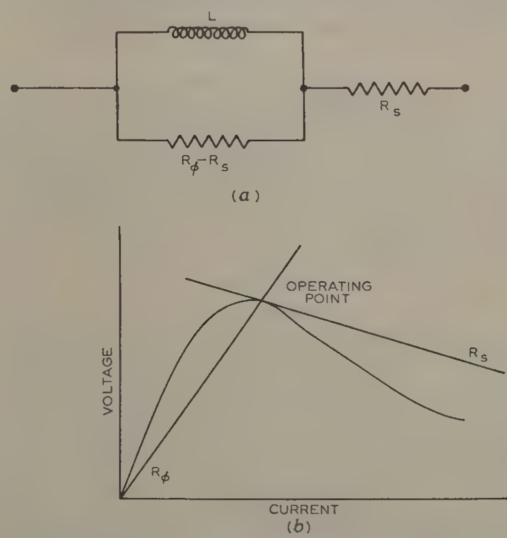


Fig. 14.—Small-signal equivalent circuit and its relationship to the V/I characteristic.

and will be negative once the current has exceeded the value at V_{max}. R_ϕ is the slope of the line from the origin to the operating point and represents the high-frequency resistance of the thermistor. The value of the inductance at a given point on the static characteristic is proportional to the time-constant of the thermistor multiplied by R_ϕ and can, in practice, be greater than 10^6 H. The presence of negative resistance and inductance are dependent on the thermistor having a negative temperature coefficient. A resistor with a positive temperature coefficient would have positive resistance and capacitance in the corresponding equivalent circuit.

By applying a direct bias to the thermistor, or by applying heater power to an indirectly-heated thermistor,[73] and by shunting the thermistor with a capacitor, simple very-low-frequency oscillators may be made.[5, 32, 74, 75] Similarly, filters[74] and modulators[5] may be designed.

(7.2) Large-Signal Operation

Large-signal transient problems can be analysed graphically[65] if the assumption is made that the temperature distributions within the thermistor under steady-state and transient conditions are the same. This is not exactly true, but in practice the errors are usually small and it would be possible, for example, to estimate with fair accuracy the delay time of a relay in series with a thermistor. Circuits with more than one temperature-sensitive element, such as a thermistor in series with a lamp, are not easy to treat analytically.

The large-signal a.c. (as opposed to the transient) problem of a thermistor in series with a resistor has been solved using a mechanical differential analyser.[65]

(7.3) The Thermistor as a Logical Element

Fundamentally a thermistor is capable of being used as a logical element in digital data-handling systems, but its slow speed and temperature dependence make its widespread use unlikely in this field. Such use would depend on its 'breakdown' if the voltage across it exceeded V_{max} and is illustrated by its use in 'fail-safe' circuits. For example, if several lamps are connected in series, each having a thermistor shunt, and one lamp fails, the voltage across its thermistor will rise above V_{max} and the thermistor will change to the high-conduction state. The thermistor's hot resistance can be chosen to equal that of a lamp.

(7.4) Noise in Thermistors

Knowledge of a thermistor's noise behaviour is incomplete. There is evidence that at low frequencies a thermistor carrying a current is not as noisy as a carbon resistor,[76] but as far as the authors are aware it is not known how noisy thermistors are at high and very-high frequencies. Attempts have been made to improve the noise performance of flake thermistors for use in infra-red bolometers.[77]

(8) CONCLUSIONS

Since its conception some twenty-five years or more ago and its realization as a practical device about ten years later, understanding of how the thermistor works and what it can do has grown enormously. Is there more to be learnt? And if so, what? With the ever-increasing knowledge of semiconductors, which has caused, and is caused by, the advent of the transistor, the possibilities of new thermistor materials and a better understanding of these materials is very evident. In applying the thermistor, use has been made of all its basic thermal and electrical properties and most of these have been analysed; there is still work to be done, but the possibilities are limited. The thermistor is certainly one of the most versatile 2-termin devices known in electrical engineering; it has proved itself t be most robust, reliable, compact and sensitive and in mar fields can compete on equal terms with other components whic are more specialized in their application.

(9) ACKNOWLEDGMENTS

Thanks are due to the Directors of Standard Telephones an Cables, Ltd., for permission to publish much of the data co tained in the paper, to Marconi Instruments, Ltd., for detai of measurements of the frequency-resistance characteristics thermistors in the v.h.f. range on which the conclusions on th subject are based, and to many of the authors' colleagues wh helped with information and advice.

(10) BIBLIOGRAPHY

(1) BECKER, J. A., GREEN, C. B., and PEARSON, G. L.: 'Properties ar Uses of Thermistors—Thermally Sensitive Resistors', *Transactio of the American I.E.E.*, 1946, **65**, p. 711.
(2) GOUDET, G., and MEULEAU, C.: 'Semiconductors: Their Theory ar Practice' (Macdonald and Evans, 1957).
(3) VERWEY, E. J. W., HAAYMAN, P. W., and ROMEYN, F. C.: 'Semico ductors with Large Negative Temperature Coefficience of Resi tance', *Philips Technical Review*, 1947–48, **9**, p. 239.
(4) BOSSON, G., GUTMAN, F., and SIMMONS, L. M.: 'A Relationsh between Resistance and Temperature of Thermistors', *Journal Applied Physics*, 1950, **21**, p. 1267.
(5) SMITH, O. J. M.: 'Thermistors. Part I. Static Characteristics' ar 'Thermistors. Part II. Dynamic Characteristics', *Review Scientific Instruments*, 1950, **21**, pp. 344 and 351.
(6) SACHSE, H. B.: 'Low Resistance Thermistors as Ultra-Cold Therm meters', *Electronic Industries*, January, 1957, **16**, p. 55.
(7) BEAKLEY, W. R.: 'The Design of Thermistor Thermometers wi Linear Calibration', *Journal of Scientific Instruments*, 1951, **2** p. 176.
(8) KEONJIAN, E., and SCHAFFNER, J. S.: 'Shaping of the Characteristi of Temperature-Sensitive Elements', *Transactions of the Americ I.E.E.*, 1954, **73**, Part I, p. 396.
(9) SOBLE, A. B.: 'Bounds for Thermistor Compensation of Resistan and Conductance', *Transactions of the Institute of Radio Engineer* 1957, **CP–4**, p. 96.
(10) COLE, K. S.: 'Thermistor Thermometer Bridge: Linearity and Sens tivity for a Range of Temperature', *Review of Scientific Instrumen* 1957, **28**, p. 326.
(11) PEARSON, G. L.: 'Thermistors, Their Characteristics and Uses', *Be Laboratories Record*, 1940, **19**, p. 106.
(12) BECK, A.: 'The Stability of Thermistors', *Journal of Scientific Instr ments*, 1956, **33**, p. 16.
(13) MULLER, R. H., and STOLTEN, H. J.: 'The Use of Thermistors Precise Measurement of Small Temperature Differences (Therm metric Determination of Molecular Weights)', *Analytical Chemistr* 1953, **25**, p. 1103.
(14) GREENHILL, E. B., and WHITEHEAD, J. R.: 'An Apparatus for Measurir Small Temperature Changes in Liquids', *Journal of Scientific Instr ments*, 1949, **26**, p. 92.
(15) JOHNSON, P.: 'The Calibration and Use of an Air-driven Ultracentr fuge', *ibid.*, 1951, **34**, p. 91.
(16) ANDERSON, J. C.: 'Temperature Measurement with Thermistors *Electronic and Radio Engineer*, 1958, **35**, p. 80.
(17) HERINGTON, E. F. G., and HANDLEY, R.: 'The Use of Thermisto for the Automatic Recording of Small Temperature Differences *Journal of Scientific Instruments*, 1948, **25**, p. 430.
(18) COUTTS, J. R. H.: 'Observation of Soil Temperatures in North Ea Scotland', Sixième Congrès de la Science du Sol, Paris, 1956.
(19) DEETER, E. L.: 'Null Temperature Bridge', *Electronics*, May, 194 **21**, p. 180.
(20) BADGLEY, F. I.: 'Response of Radiosonde Thermistors', *Review Scientific Instruments*, 1957, **28**, p. 1079.
(21) HALES, W. B.: 'Thermistors as Instruments in Thermometry an Anemometry', *Bulletin of the American Meteorological Societ* 1948, **29**, p. 494.
(22) HAUCK, V. D., COSBY, J. R., and DEMBER, A. B.: 'Radiosonde Tel metering Systems', *Electronics*, May, 1946, **19**, p. 120.
(23) BATES, S.: 'Bearing Temperature Control', *Technical Journal of th Brush Group*, 1954, **2**, p. 2.
(24) KROG, J.: 'Thermistor Hypodermic Needle for Subcutaneous Tem perature Measurement', *Review of Scientific Instruments*, 1956, **2** p. 408.
(25) PATTEE, T. R.: 'Temperature Compensation of Instruments wit Negative Coefficient Resistors', *Electronics*, August, 1943, **16**, p. 10
(26) GENERAL ELECTRIC CO. (METALLURGICAL PRODUCTS DEPT.): ' Thermistor Selecting Shortcut', *Control Engineering*, 1957, **4**, p. 7
(27) GUENÉ, P.: 'The Use of Thermistors for the Compensation of Dri in Self-Oscillating Circuits', *Annales de Radioélectricité*, 1956, **11** p. 317.

(28) 'Temperature-Compensated Crystal Oscillator', *Electronic and Radio Engineer*, 1958, **35**, p. 314; also *Wireless World*, 1958, **64**, p. 441.
(29) BOLLMAN, J. H.: 'A Pilot-Channel Regulator for the K–1 Carrier System', *Bell Laboratories Record*, 1942, **20**, p. 258.
(30) JOHNSON, J. C.: 'Thermistor Technics', *Electronic Industries*, August, 1945, **4**, p. 74.
(31) WHEELER, A. J.: 'Thermistors Compensate Transistor Amplifiers', *Electronics*, January, 1957, **30**, p. 169.
(32) HOWES, T. W.: 'Characteristics and Applications of Thermally Sensitive Resistors or Thermistors', *Electrical Communication*, 1955, **32**, p. 98.
(33) DICKINSON, C. J.: 'Thermistor Continuous Temperature Control for Biological Research', *Electronic Engineering*, 1949, **21**, p. 408.
(34) TAYLOR, A. H.: 'Thermistor Thermostats', *Electronics*, July, 1950, **23**, p. 154.
(35) VODDEN, H. A.: 'A Thermostat Control Unit', *Journal of the Society of Chemical Industries*, 1950, **69**, p. 51.
(36) British Patent No. 10965/56.
(37) AMBROSE, D., and COLLERSON, R. R.: 'A Thermal Conductivity Gauge for Use in Gas-Liquid Partition Chromatography', *Journal of Scientific Instruments*, 1955, **32**, p. 323.
(38) DAVIS, A. D., and HOWARD, G. A.: 'The Use of Thermistor Detectors in Gas Chromatography', *Journal of Applied Chemistry*, 1958, **8**, p. 183.
(39) WALKER, R. E., and WEATENBERG, A. A.: 'Precision Thermal-Conductivity Gas Analyser Using Thermistors', *Review of Scientific Instruments*, 1957, **28**, p. 789.
(40) VARICAK, M.: 'Investigation of the Use of Thermistors for the Measurement of Low Pressure', *Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences*, 1956, **243**, p. 893.
(41) DAVIS, F. E.: 'Vacuum Furnace Control', *Electronics*, May, 1958, **21**, p. 81.
(42) PENMAN, H. L., and LONG, I.: 'A Portable Thermistor Bridge for Micro-meteorology among Growing Crops', *Journal of Scientific Instruments*, 1949, **26**, p. 77.
(43) CHUDNOVSKI, A. F.: 'The Use of Semiconductors in Investigating Natural Phenomena', *Zhurnal Tekhnicheskoĭ Fiziki*, 1955, **25**, p. 2122.
(44) WARD, W. H.: 'Two Portable Thermistor Radiometers', *Journal of Scientific Instruments*, 1957, **34**, p. 317.
(45) 'Thermal Detectors', *Electronic Industries*, September, 1946, **5**, p. 87.
(46) WIDDIS, F. C.: 'The Indirectly Heated Thermistor as a Precise A.C.–D.C. Transfer Device', *Proceedings I.E.E.*, Paper No. 1989 M, February, 1956 (**103** B, p. 693).
(47) SCHRADER, H. J.: 'A.C. Compensation Measurement with Indirectly Heated N.T.C. Resistors', *Elektrotechnische Zeitschrift*, 1952, **73** A, p. 547.
(48) WOOD, H. B.: 'An R.M.S. Milliammeter of Novel Design for the Measurement of Current from Zero to Video Frequencies', *Journal of Scientific Instruments*, 1954, **31**, p. 124.
(49) SOLDI, M.: 'Applications of Thermistors to Bridges for Measuring Low Conductance at High Frequencies', *Alta Frequenza*, 1952, **21**, p. 243.
(50) WALKER, J.: 'Thermistors at High Frequencies', *Wireless Engineer*, 1947, **24**, p. 28.
(51) LANE, J. A.: 'The Measurement of Power at a Wavelength of 3 cm by Thermistors and Bolometers', *Proceedings I.E.E.*, Paper No. 1981 R, November, 1955 (**102** B, p. 819).
(52) RIECH, H., and PANNIGER, F.: 'A New Thermistor Power Measuring Head for $\lambda = 9 - 20$ cm', *Nachrichtentechnik*, 1957, **7**, p. 101.
(53) STREET, R.: 'The Absolute Measurement of Low Power at 3 000 Mc/s', *Proceedings I.E.E.*, Paper No. 838 M, June, 1949 (**96**, Part II, p. 391).
(54) BATH, C. C., and GOLDBERG, H.: 'Self Balancing Thermistor Bridge', Proceedings of the National Electronics Conference, 1947, **3**, p. 45.
(55) COLLARD, J., NICOLL, G. R., and LINES, A. W.: 'Discrepancies in the Measurement of Microwave Power at Wavelengths below 3 cm', *Proceedings of the Physical Society*, 1950, **63** B, p. 215.
(56) PEARSON, R. M., and BENSON, F. A.: 'The Design of a Direct-Reading Thermistor Bridge', *Electronic Engineering*, 1953, **25**, p. 51.
(57) MEACHAM, L. A.: 'The Bridge-Stabilized Oscillator', *Proceedings of the Institute of Radio Engineers*, 1938, **26**, p. 1278.
(58) SHEPHERD, W. G., and WISE, R. O.: 'Variable-Frequency Bridge-Type Frequency-Stabilized Oscillators', *ibid.*, 1943, **31**, p. 256.
(59) WILLIAMSON, R.: 'Wide Range Audio Oscillator', *Wireless World*, 1956, **62**, p. 508.
(60) FLEMING, L.: 'Thermistor Regulated Low Frequency Oscillator', *Electronics*, October, 1946, **19**, p. 97.
(61) BLEUZE, J.: 'Les Propriétés et les Applications des Thermistances', *L'Onde Electrique*, 1953, **33**, pp. 497 and 578.
(62) HARDIE, A. M.: 'A Thermistor Device as Automatic Gain Control in Lamp Photocell Transducer Systems', *Journal of Scientific Instruments*, 1957, **34**, p. 58.
(63) WELLER, J. A.: 'A Volume Limiter for Leased-Line Service', *Bell Laboratories Record*, 1945, **23**, p. 72.
(64) PATCHETT, G. N.: 'Precision A.C. Voltage Stabilisers', *Electronic Engineering*, 1950, **22**, pp. 371, 424, 470 and 499.
(65) EKELOF, S., BJORK, N., and DAVIDSON, R.: 'Large Signal Behaviour of Directly Heated Thermistors', *Chalmers Tekniska Hogskolas Handingar*, 1957, No. 185.
(66) WHITEHEAD, R. C.: 'A Programme Fading Circuit', *B.B.C. Quarterly*, 1953/54, **8**, No. 4 (Winter).
(67) LINAU, H., and SEIFERT, I.: 'Relay Time-Delay Circuits with N.T.C. Resistors', *Matronics*, 1957, **12**, p. 216.
(68) GANO, J. J., and SANDY, G. F.: 'Thermistors for the Gradual Application of Heater Voltage to Thermionic Tubes', *Transactions of the Institute of Radio Engineers*, 1958, **EC–7**, p. 61.
(69) SILLARS, R. W.: 'Materials and Devices of Falling Resistance—Temperature Characteristic', *Journal of Scientific Instruments*, 1942, **19**, p. 81.
(70) CAMERON, C. T., and HUMPHRIES, W. A.: 'Shared Service', *Post Office Electrical Engineers' Journal*, 1948, **41**, p. 139.
(71) EKELOF, S., and KIHLBERG, G.: 'Theory of the Thermistor as an Electric Circuit Element', *Chalmers Tekniska Hogskolas Handlingar*, 1954, No. 142.
(72) BJORK, N., and DAVIDSON, R.: 'Small Signal Behaviour of Directly Heated Thermistors', *ibid.*, 1955, No. 169.
(73) ROSENBERG, W.: 'Thermistors', *Electronic Engineering*, 1947, **19**, p. 185.
(74) CANDY, C. J. N.: 'The Specification of the Properties of the Thermistor as a Circuit Element in Very-Low-Frequency Systems', *Proceedings I.E.E.*, Paper No. 1751 M, December, 1954 (**103** B, p. 398).
(75) STONE, J. E.: 'An Ultra-Low Frequency Oscillator', *Electronics*, January, 1950, **23**, p. 94.
(76) VAN VLIET, K. M., VAN LEEUWEN, C. J., BLOK, J., and RIS, C.: 'Measurements on Current Noise in Carbon Resistors and Thermistors', *Physica*, 1954, **20**, p. 481.
(77) 'Electron Microscopic Study of Sintering in Thermistor Flakes', *Bell Laboratories Record*, 1952, **30**, p. 285.
(78) BURGESS, R. E.: 'The A.C. Admittance of Temperature Dependent Circuit Elements', *Proceeding of the Physical Society*, 1955, **68** B, p. 766.
(79) HYDE, F. J.: 'Reactive Effects in Thermistors at Very Low Frequencies', *British Communication and Electronics*, 1957, **4**, p. 16.
(80) HYDE, F. J.: 'The Impedance of a Thermistor at Low Frequencies', *Journal of Electronics*, 1955, **1**, p. 303.

DISCUSSION BEFORE THE MEASUREMENT AND CONTROL SECTION, 5TH APRIL, AND THE NORTH-WESTERN MEASUREMENT AND CONTROL GROUP, AT MANCHESTER, 16TH FEBRUARY, 1960

Dr. F. C. Widdis: Some years ago I made a study of the use of the indirectly-heated thermistor as a transfer device for the precise measurement of alternating currents (described in Reference 46). To compensate for ambient temperature changes it was necessary to use two of these thermistors in a bridge arrangement, and the resulting drift was equivalent to $1\cdot5$ parts in 10^4 of the measured current over a period of three days. Good temperature compensation in such a bridge arrangement is only achieved if the two thermistors have the same temperature coefficient, $\alpha = -B/T^2$. If the two thermistors have the same value of B this condition is easily satisfied by keeping them at the same temperature, T; what tolerances might be expected on the value of B for a given batch of thermistors?

I was surprised at the high degree of reproducibility shown by the thermistors in this work. I had no difficulty in repeating current measurements with a precision of a few parts in 10^5, equivalent to a temperature stability of $\pm0\cdot005°$ C in 60° C rise. I gained the (unverified) impression, however, that high-resistance thermistors are more reproducible than low-resistance ones, and would appreciate the authors' comments.

I have managed to show that the indirectly-heated thermistor used as a transfer device has an error much less than 1 part in 10^4 over a frequency range of $0\cdot5$ c/s–50 kc/s. The authors have stated that the upper frequency limit of this device is determined by the inductance of the heater—actually the self-capacitance of the heater is of major importance.

The thermistor is a simple and very sensitive alternative to the thermocouple for measuring temperature in confined spaces. My own experience has shown that they are very reproducible over limited temperature ranges for short periods, and although the results given in Fig. 9 are promising, additional work on their behaviour over extended temperature ranges and long periods would be of considerable value.

Dr. A. J. Maddock: Can the authors give more details of the improved temperature stability of modern thermistors? Anyone desirous of measuring temperature wants to know first of

all how stable these devices are on a long-term basis, and although a few papers have appeared previously giving results of this nature, they have all referred to the older type of thermistor. It would also be helpful if the firms which manufacture these devices gave this data in their trade literature.

The authors have quoted the results of a year's run on 50 elements of a single type in Fig. 9. Are similar results obtained on other types? The results they give are virtually shelf or storage stability: they have tested the thermistors and only cycled them once prior to each test up to 100° C. Have they any results which indicate the stability of these devices when they are continually in use at some elevated temperature—perhaps 100° C?

Several circuits for linearizing the response of thermistors exist in the literature. There are also circuits available for matching two thermistors one to another: is this necessary with modern thermistors if one is trying to obtain balanced characteristics?

Mr. J. D. Brownlie: We have been concerned at Northampton College to find a general method of predicting the waveforms of large alternating voltages, currents and temperatures in a temperature-sensitive element such as the thermistor.

A simple very-low-frequency oscillator circuit can be formed by paralleling a suitably biased thermistor with a large capacitor (Section 7.1) as in Fig. A.

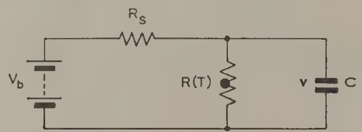


Fig. A.—Oscillator circuit.

The conditions for starting and at very small amplitudes can be estimated well enough by using the small-signal equivalent circuit of the thermistor. As soon as the amplitudes start building up, however, the incremental theory no longer applies and the waveforms become increasingly non-sinusoidal.

If the assumption is made that the whole of the thermistor element is at a uniform temperature at any given time, the operation of the circuit can be described by the two simultaneous differential equations

$$H\frac{dT}{dt} = \frac{v^2}{R(T)} - k(T - T_A)$$

$$C\frac{dv}{dt} = \frac{V_b}{R_s} - \frac{v}{R_s} - \frac{v}{R(T)}$$

where H is the incremental heat capacity and k is the incremental dissipation constant.

The first gives the instantaneous rate of heat storage and the second, the rate of increase of charge on the capacitor. For any but incremental changes in the variables, the equations can only be solved numerically.

We have recently finished programming a digital computer to give step-by-step solutions of voltage and temperature using the Runge–Kutta method. The process can be made very accurate by choosing a small time interval. Thermistor resistances at the required temperatures are found by interpolation in a Table of experimental values.

The computed waveform of voltage appears to be very similar to the actual waveform. When a proper comparison has been made the results will serve to test the validity of the theory. In addition, the computation will enable us to find very quickly the effects on the circuit of changes in, e.g., ambient temperature, heat capacity and thermal dissipation. It is hoped that, later similar computation can be applied to a circuit in which a resistor with a positive temperature coefficient supplies the necessary large capacitive reactance.

This method (first suggested to me by Dr. House) offers some clear advantages over analogue computing methods[65] in the ease of recording, the accuracy and the facility of exchanging accuracy for speed of calculation. We think, therefore, that this oscillator circuit has provided an illustration of a good general method of solving non-linear circuits as well as being an interesting application of 'thermal reactance'.

Mr. A. C. Lynch: If the current/voltage response of a device is unusual, it may be impossible to represent it by an equivalent circuit, and I think this is true of a thermistor. The large inductance referred to in Section 7.1 describes its response to direct current, but not to alternating current, to which the thermistor appears as a resistance varying with time. If the proposed equivalent circuit were really valid, it could be used to deduce the inductance observed at high frequencies and referred to in Section 6.1; but of course the two inductances are quite unrelated. That observed at high frequencies is to be expected; for it is generally true that any low resistance (below, say, 100 ohms) will appear to be inductive; the self-capacitance which tends to compensate for the inductance is not large enough, and often cannot be made large enough, to compensate for it entirely.

Dr. A. K. Jonscher: The thermistor being essentially a compacted granular element, it would, at least in principle, be expected to exhibit surface barrier effects. It is not clear to me what is meant by the statement in Section 1 that conduction is purely electronic, and that current is proportional to applied voltage. One might infer that barriers do not come into consideration since they are normally strongly non-linear. There are two possibilities: either the barriers are ineffective, or the voltage drop per barrier is sufficiently small in the regime under consideration to give linear behaviour. According to whether barriers are or are not effective, the resistance of the compacted body would bear some relation to the expected resistance of single-crystal material under those conditions. Is anything known about this?

If the barriers were responsible for greater voltage-dependence on resistance, could some contribution of this type be expected in Fig. 4? Presumably the strong non-linearity is largely thermal in nature, but is there any contribution by barriers there?

As regards dependence of resistance upon temperature, it is known that mobility in ionic crystals can be exponential in temperature rather than follow a power law. Is anything known about the relative influence of mobility and carrier density variation on the temperature-dependence of thermistors? I believe that at one stage silicon-carbide compacted bodies were used as thermistors. There appears to be no reason why only ionic compounds should be suitable for this purpose; will the authors comment on the relative merits of silicon-carbide compacted bodies as against the ionic kind?

Dr. J. Evans: Single-crystal semiconductors can also be used as thermistors. The resistivity/temperature curve for such a semiconductor is shown in Fig. B. The thermistors we have been talking about are represented by the dotted line, and a copper-wire resistor by the broken line. One can make a device with a positive temperature coefficient which, in the case of silicon, is about 0·7% per deg C due to mobility change. On the other hand, one can use the intrinsic portion which has a negative coefficient of several per cent per degree. I think that both can probably be used, and perhaps we shall see an extension into this field as the art progresses.

On the question of crystallinity, a similar effect can also be

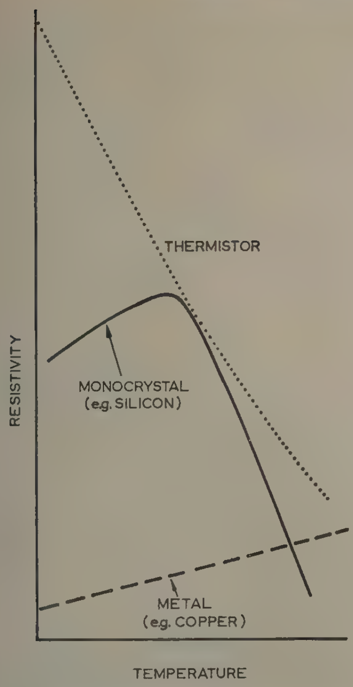


Fig. B.—Resistivity/temperature characteristics.

found in silicon. For example, if the silicon is made by a certain process* it comes out in the form of polycrystalline agglomerates. If that crystal is melted and then re-frozen, it will grow into larger crystallites and the resistivity/temperature plot for this material is identical with the plot for a single-crystal material. I think one can conclude that in a number of cases the barriers themselves seem to be inactive and the properties are those of the individual grains of the material.

Mr. M. A. Snelling: I also am interested in the stability of thermistors, particularly when they are used for thermal measurements. Are ageing characteristics of the type shown in Fig. 9 attributed to practical features of manufactured thermistor units, or are they believed to be an inherent property of the thermistor material? Can the authors give some indication of the subsequent changes in value which would have occurred to the units used in obtaining Fig. 9 if the measurements had been continued for a further 50 weeks? What steps should thermistor users take to minimize subsequent drift occurring after the ageing process is complete?

I note that conduction in thermistors is believed to be purely electronic. It presumably follows that within reasonable ranges of power and frequency either a.c. or d.c. operation of thermistors will give similar stability. Is this supposition confirmed in practice?

Mr. W. H. P. Leslie: In the past it has been very difficult to obtain information from any manufacturer on the performance to be expected from thermistors for temperature measurement. The stock answer has been 'test for yourself' with the implication that, since the maker has no information on stability with life, one must buy one big batch, test some over a year or so, and use only the tested batch for future work.

Can the authors indicate whether they hold any hope of measuring to 0·01° C or better for practical conditions of temperature cycling in the range 0–100° C with calibration once, say, in three months?

Most attempts at precision temperature measurement using resistance thermometers have been based on the use of pure materials and strain-free mountings. Thermistors appear, from the paper, to be made from a bit of this and a bit of that with some binding agent, all mixed together and dabbed on wires before cooking. Will the authors agree that this may be a reason for the changes in calibration in practical use, and also that the different coefficients of expansion of wire and bead must introduce temperature-dependent strain and thus cause ageing?

The paper claims that thermistors have been used to measure temperature differences of 0·002° C over short periods. Can one assume that for this type of measurement it is necessary to use one thermistor exposed successively to the two temperatures and that the use of two thermistors at two points to indicate instantaneous temperature difference could not yield this accuracy over any useful range of ambient temperatures?

Mr. H. R. Westaway: Being engaged in the manufacture of temperature measuring instruments I have been very interested in the practicability of using thermistors as alternatives to wire-wound resistance-thermometer elements and have carried out extensive investigations into their stability. The results have been largely in line with those given in the paper, and my conclusion is therefore that they cannot compare with the conventional nickel or platinum elements in this respect.

The thermistor has the advantages of small size, high speed of response and sensitivity, but a disadvantage is the low standard of interchangeability between elements of a given type.

The minimum tolerance generally available on basic resistance values is $\pm 5\%$ and, in addition, there is an unspecified tolerance on the α-value. Do the authors consider that there is any prospect of closer tolerances being forthcoming with improved methods of manufacture?

It is essential that, for industrial temperature measurement, particularly in multi-point installations, arrangements be made for standardizing the thermistors. This necessitates a parallel resistor to standardize the resistance change over a given temperature interval and a series resistor to bring the resistance up to the nominal value at a given temperature, giving a relatively cumbersome assembly at the measuring point by comparison with the thermistor itself.

Another limitation which has been very evident, particularly with regard to the bead type, is that of the maximum current which can be passed through the thermistor without causing prohibitive errors due to self-heating. When a deflectional-type measuring instrument is being used, this aspect necessitates the use of a moving-coil system having a relatively high sensitivity.

Dr. H. J. Goldsmid: I suggest that the reason why thermistors are as stable as they appear to be is that thermistor materials have large carrier concentrations so that very little change of the conductivity results from changes in impurity concentration. On the other hand, with a material such as pure silicon, a very small change of impurity concentration affects the carrier concentration or mobility. I do not, therefore, think that better stability would be obtained with silicon or germanium.

The upper temperature limit shown in the authors' curves is a few hundred degrees Celsius. Can thermistors be used at much higher temperatures, say up to 1 000° C, using, of course, different materials?

Messrs. I. C. Hutcheon and **D. N. Harrison** (*communicated*): The use of thermistors as controllable resistors, rather than as temperature-sensing devices, possibly merits rather more attention than it has received in the paper.

Two indirectly-heated thermistors, for example, can be controlled differentially by a demodulated a.c. signal so that the

* GIBSON, A. F. (Editor): 'Progress in Semiconductors, Vol. 3' (Heywood, 1958), p. 42—WILSON, J. M.: 'The Chemical Purification of Germanium and Silicon'.

resistance of one rises while that of the other falls by about the same amount. If the thermistors are connected together, their common point can be driven from one end to the other of the total resistance, and the arrangement shown in Fig. C is similar to an a.c. amplifier driving a reversible servo-motor coupled to the contact on a slide wire. The main differences are that the position of the 'contact' on the thermistor potentiometer is not known, the system is not accurately linear, and the response is rather slow.

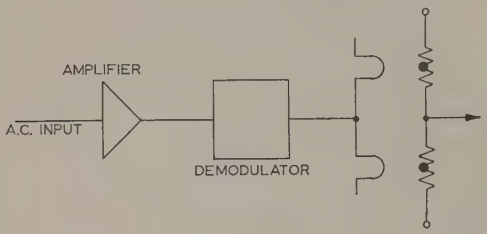


Fig. C.—Thermistor potentiometer.

Nevertheless such a device,* which can be built cheaply using only low-power transistors, thermistors and other solid-state components, has quite wide application. Connected in a subsidiary negative-feedback loop around an a.c. servo, for example, as shown in Fig. D, it suppresses any quadrature components which may be present in the input or the main feedback carrier signals, so eliminating a common cause of error and preventing amplifier saturation. Two such devices, operating in quadrature with one another, can form the basis of an accurate phase-meter.

In another application, the potentiometer can be shared in time between two d.c. circuits by the use of synchronous switches, so permitting the accurate multiplication of two slowly varying inputs.

Mr. D. Williamson (*at Manchester*): Would it be feasible to manufacture thermistors in the form of pairs with matched temperature coefficients, electrically insulated from each other but in good thermal contact, and indirectly heated by a common heater? By varying the heater current until the resistance of one of the thermistor elements was of a desired value, the resistanc[e] of the other element would automatically follow. This elemen[t] could then be used in a separate circuit. Suitable closed-loo[p] technique, whereby the heater current is supplied by an erro[r] operated amplifier, should make it possible to obtain an accurat[e] electronically-controlled variable resistance, which should hav[e] many applications, e.g. simple analogue multipliers, variable-gai[n] amplifiers.

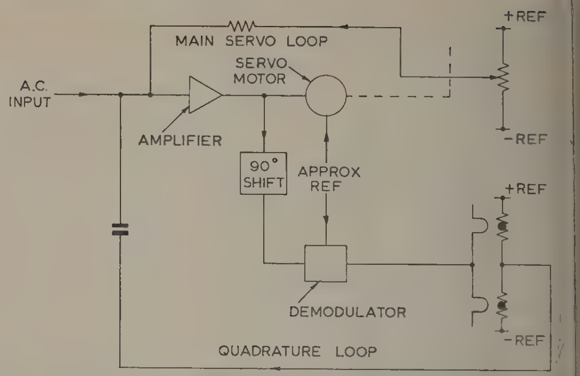


Fig. D.—Thermistor potentiometer used for quadrature suppression.

Would not a composite unit of the type described be preferabl[e] to using two separate indirectly-heated thermistors with commo[n] heater current, as, in so far as the close thermal contact ensure[d] that both elements were at the same temperature, errors woul[d] not be introduced by different self-heating effects in the tw[o] thermistor elements when different currents were flowin[g] through them?

Mr. P. J. Stratfold (*at Manchester*): Will thermistors with[stand] immersion for long periods in liquids?

Mr. F. Crowther (*at Manchester*): When used for temperatur[e] measurements, it would appear that the accuracy of a thermisto[r] is dependent on the current circulated and the self-generate[d] heat. A thermocouple would appear simpler to use as it can b[e] arranged to give results independent of circuit resistance. Ho[w] do the two methods of measurement compare?

THE AUTHORS' REPLY TO THE ABOVE DISCUSSION

Dr. R. W. A. Scarr and **Mr. R. A. Setterington** (*in reply*): Dr. Widdis considers that the self-capacitance of the heater winding of an indirectly-heated thermistor may be of major importance. The relative effects of the series inductance and the shunt capacitance will depend on the source impedance of the signal applied to the heater. The effect of series inductance will be greatest with a constant voltage supply and the effect of shunt capacitance will be greatest with a constant current supply. Our measurements show that the inductive reactance of the 100-ohm heater of a type B thermistor is 5μH at 10 Mc/s; of course, the reactance will become capacitive above the self-resonant frequency.

In general, thermistors made from materials having lower specific resistances are less tolerant of operation at high temperatures; when operated well within their temperature range we would expect them to be as reproducible as high-resistance units.

No detailed results have been obtained for stability over a long period, although Dr. Maddock will be pleased to hear that we are starting new tests to enable us to give the details he requires. It is generally considered that cycling thermistors is more stringent than operation at steady temperature; thus, results obtained from running thermistors at 100° C continuousl[y] would be no worse than those of Fig. 9.

In reply to Mr. Lynch, the small series inductance associate[d] with leads could be added in series with R_s in Fig. 14(*a*) to giv[e] a more complete equivalent circuit. However, the applicatio[n] of the circuit of Fig. 14(*a*) will be almost entirely confined t[o] frequencies in the audible range or lower, where lead inductanc[e] is insignificant. The equivalent circuit is a perfectly vali[d] representation of the thermistor's behaviour to a low-frequenc[y] small-signal alternating current.

Dr. Jonscher asks what is meant by electronic conduction[:] the term is used in contradistinction from ionic conduction[.] Nothing is known about single crystals of the materials tha[t] are commonly used in thermistor manufacture. The non[-]linearity of the characteristic shown in Fig. 4 is entirely thermal[;] any non-linearities due to barriers, if they exist at all, are of [a] very low order indeed. We do not know of any theories o[f] conduction in the metallic oxides that could be used to answe[r] the remainder of Dr. Jonscher's questions.

In reply to Mr. Snelling, ageing as shown in Fig. 9 is assume[d] to be due to chemical and physical changes in the materials[.] It happens in all types of thermistors but is considerably reduce[d]

* Hutcheon, I. C., and Harrison, D. N.: 'A Transistor Quadrature Suppressor for A.C. Servo Systems', *Proceedings I.E.E.*, Paper No. 3134 M, January, 1960 (107 B, p. 73).

in those where the active material is sealed into solid glass. Generally, resistance changes due to ageing decrease with time. It makes no difference whether thermistors are operated with a.c. or d.c. measuring equipment.

In reply to Mr. Leslie, the materials from which thermistors are made are all carefully chosen to have coefficients of expansion similar to those of the lead wire and the glass. We believe that the match obtaining in thermistor units is better than that in the normal resistance thermometer.

There is no reason why it should be much more difficult to measure a small temperature difference using two thermistors than it is with one. Differences in B-values are generally so small that they have a second-order effect. The results described in Reference 14 do in fact involve the characteristics of a pair of thermistors.

We are experimenting with materials for operation up to 1 000° C to solve Dr. Goldsmid's problems. The units are not yet sufficiently developed to allow them to be used commercially.

We would like to thank Messrs. Hutcheon and Harrison for providing a reference to their paper on a subject that inevitably could not receive the comprehensive treatment that it deserves.

Mr. Williamson's double-bead unit is not a good practical proposition for several reasons. It is difficult to match a pair of beads for slope and resistance before ageing. When this has been done it is even more difficult to obtain exactly the same degree of thermal coupling between the heater and each of the beads. Would it not be possible to measure the resistance of a single bead by two separate currents, say direct and at 1 kc/s and thus eliminate all errors due to using two different beads?

The answer to Mr. Stratfold's question is yes, provided that the thermistor is protected by glass (i.e. it is not the block or rod type) and that the liquid concerned does not react chemically with the glass.

Mr. Crowther will find part of the answer to his question in the paragraph under Fig. 9. The fact that a thermistor is heated by the current that is needed to determine its resistance is of no significance when it is working in a specified medium. When working in a variety of environments it is important to keep self-heating to a minimum, but in practice it is quite possible to use most bead thermistors in a bridge circuit with, say, a 50μA or 100μA meter as detector.

It is convenient to reply to the questions involving tolerances and stability together. We must emphasize that our experiences are limited to one manufacturer's products.

Thermistors from the same batch often have B-values within $\pm\frac{1}{2}\%$ although this can rise to $\pm 2\frac{1}{2}\%$. The spread can be as wide as $\pm 10\%$ when thermistors are made from different batches of raw materials. In the case of thermistors used for temperature measurement, pairs matched to within 2% at 20° C are usually available, whilst larger groups or tighter tolerances can be selected as required.

Although Mr. Leslie regrets the reply 'test for yourself' and draws an unfair implication from it, we are often forced into giving a similar reply because we have no knowledge of the particular set of conditions under which the customer intends to use the thermistor. Apart from Fig. 9, which gives a general picture of the stability which can be expected, several of the references give results under other conditions.

AN ANALOGY BETWEEN NON-LINEAR RESISTIVE AND LINEAR A.C. NETWORKS

By D. Q. Mayne, M.Sc.(Eng.), Graduate.

(*Communication received 7th April*, 1960.)

To every a.c. network there corresponds a non-linear d.c. network in which the relations between the various non-linear resistance elements are determined by the a.c. network. In certain simple cases the non-linear resistance elements become duals and the analogy mentioned by Cherry* in a recent paper results.

Some of the non-linear resistive networks described in Cherry's paper are analogous to Zobel's constant-resistance phase equalizers; in other cases the analogy fails. In seeking the physical basis of the analogy it would be convenient to have a rectangle-diagram representation of the analogous a.c. network. Instantaneous values are used to describe the non-linear resistive network, but complex quantities are required to describe an a.c. network. Complex quantities cannot be used directly in a rectangle diagram. However, the basis of the rectangle diagram is Kirchhoff's laws and these apply separately to the in-phase and quadrature components of the alternating voltages and currents. Thus a rectangle diagram may be constructed for an a.c. network using as the sides of the rectangle V_p and I_p, V_q and I_q, V_p and I_q, or V_q and I_p, where the suffixes p and q denote in-phase and quadrature components with respect to a reference angle. Since, for an a.c. network complex power is given by $P = V\,.\,I^* = V_p\,.\,I_p + V_q\,.\,I_q + j(V_q\,.\,I_p - V_p\,.\,I_q)$, where $V = V_p + jV_q$ and $I = I_p + jI_q$, the total area of any of these rectangle diagrams is no longer equal to the total active power.

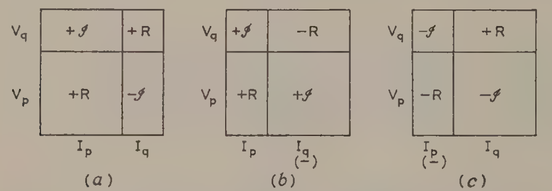


Fig. 1.—Rectangle-diagram representation for a.c. networks.

(*a*) Pure resistance.
(*b*) Pure inductance.
(*c*) Pure capacitance.

Fig. 1(*a*) shows the four rectangle diagrams corresponding to the four product terms in the expression for P, for a pure resistance. The phase angle, ϕ, of V is given by $\tan\phi = k = V_q/V_p$. For pure resistance $I_q/I_p = k$. Active power is denoted by $\mathscr{R}$ and reactive power by $\mathscr{I}$, the sign being shown in the diagram. For pure resistance the total reactive power is zero and the total active power is $(1+k^2)V_p^2/R$. In many cases the V_pI_p rectangle is most useful. The slope of this rectangle is R, the value of the resistance, and its area is V_p^2/R.

The corresponding rectangles for pure inductance are shown in Fig. 1(*b*). Here I_q is negative and $I_p/I_q = -k$. The total reactive power is $(1+k^2)V_p^2/\omega L$ and the total active power is zero. The slope of the V_pI_p rectangle is $\omega L/k$ and its area is $kV_p^2/\omega L$. The slope varies with frequency and the V_pI_p rectangle for inductance is therefore analogous to the *vi* rectangle for non-linear resistance. A V_pI_q rectangle diagram would be more 'natural' for an inductor, the slope being $\mp\omega L$ and the area $V_p^2/\omega L$.

Fig. 1(*c*) illustrates the behaviour of pure capacitance. Here I_p is negative if V_p and V_q are positive and $I_p/I_q = -k$. The total reactive power is $-(1+k^2)V_p^2\omega C$ and the total active power is zero. The slope of the V_pI_p rectangle is $-1/k\omega C$ and its area $-kV_p^2\omega C$.

The slope of the 'non-linear' L and C elements depends on k and can quite easily be negative; this results in overlapping of rectangles. The range of possible 'non-linear' characteristics is large and it is only in certain relatively simple cases that the slopes of the L and C elements become reciprocal and therefore analogous to the dual non-linear resistors employed by Cherry.

This reciprocal relation does exist for the bridged-T structure considered by Cherry. This structure is shown in Fig. 2 together with the V_pI_p rectangle diagram and associated vector diagrams. Z_1 is here chosen as $j\omega L$ and Z_2 as $1/j\omega C$. The inverse relation

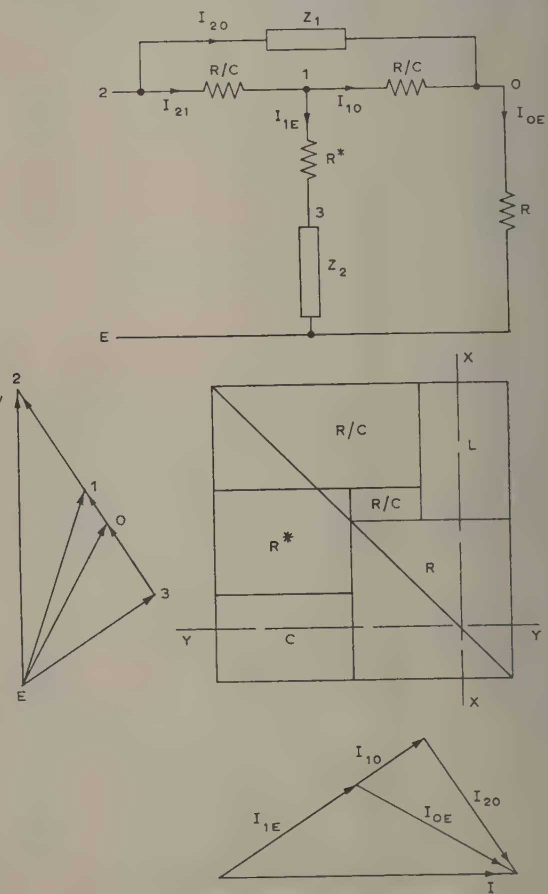


Fig. 2.—Bridged-T constant-resistance section.

* Cherry, E. C.: 'Classes of 4-Pole Networks having Non-Linear Transfer Characteristics but Linear Impedance', *Proceedings I.E.E.*, Paper No. 3114 E, January, 1960 (107 B, p. 26).

Mr. Mayne is at the Imperial College of Science and Technology, University of London.

$Z_1Z_2 = R^2$ reduces to $L/C = R^2$. The diagram is normalized so that resistance R is represented as a square. The section XX gives the equation $V_p = (V_R)_p + (V_L)_p$. This is the projection on the reference axis of the equation $V = V_R + V_L$. The section YY results in a dual equation $I = I_R + I_C$ since R and R, and ωL and $1/\omega C$, form inverse pairs. Since $V = I \, . \, R$, $V_L = I_C \, . \, R = V_C \, . \, j\omega CR$. Therefore V_L and V_C are at right angles and $k_L \, . \, k_C = 1$. The slope of the element L in the normalized rectangle diagram is $\omega L/k_L R$, and that of the element C is $1/k_C \omega CR$. The product of the slopes is unity, that is the elements are reciprocal and the network is analogous to the non-linear resistance network in which Z_1 and Z_2 are replaced by dual non-linear resistors R_1 and R_2 where $R_1 \, . \, R_2 = R^2$. This non-linear network will have a constant input resistance equal to $V_p/I_p = V/I = R$, and R^* will have the same value as for the a.c. network.

It seems that this reciprocal relation occurs whenever sections XX and YY can be constructed to give a pair of dual equations including all the reactive elements in the network. This will be the case for lattice and bridged-T structures in which Z_1 and Z_2 are inverse. The dual equations are $V = V_R + V_{Z1}$ and $I = I_R + I_{Z2}$. (For the lattice, $2V_{Z1}$ must be substituted for V_{Z1} and $2I_{Z2}$ for I_{Z2} but the reasoning is unaltered.) $V = I \, . \, R$ and $V_{Z1} = I_{Z2} \, . \, R$. If V_{Z1} makes an angle θ with the reference axis and if $\angle Z_1 = \phi$, the slope of the V_pI_p rectangle of the element Z_1 is $(V_{Z1})_p/(I_{Z1})_p = |Z_1| \, . \cos\theta/\cos(\theta - \phi)$. Similarly the slope of the rectangle for Z_2 is $(V_{Z2})_p/(I_{Z2})_p = |Z_2| \, . \cos(\theta - \phi)/\cos\theta$ as $\angle Z_2 = -\phi$. The product of the slopes is $|Z_1| \, . \, |Z_2| = R^2$ and the elements are reciprocal.

The reciprocal relation disappears, in general, in the degenerate T- and π-structures. This is illustrated in the T-structure of Fig. 3 with its associated V_pI_p rectangle diagram. Here $Z_1 = j\omega L$ and $Z_2 = 1/j\omega C$. For the particular frequency shown the 'non-linear' element $L/2$ in parallel with R adjoining the output resistance has a negative slope. A non-linear resistance network having constant input resistance can be constructed if the non-linear elements have values related to the slopes of the reactive elements in the V_pI_p rectangle diagram. These slopes are not reciprocal. The variation of the slopes with ω can be obtained if $V(\omega)$ is calculated for each element.

To every a.c. network there corresponds a non-linear resistive network. To obtain the non-linear network, resistors are left unchanged in the a.c. network but reactors are replaced by non-linear resistors. The values of the non-linear resistors for a particular input voltage are equal to the slopes of the V_pI_p rectangles of the reactors for a particular value of ω. To complete the analogy, therefore, an arbitrary relation may be established between the input voltage v of the non-linear network and ω, e.g. $v = a \, . \, \omega$. The transfer function of any known a.c. network expressed as the ratio of the input and output in-phase voltages as a function of ω is equal to the transfer function of the analogous non-linear network expressed as a function of v if we put $v = \omega$ and if the individual non-linear elements can be realized. By putting $v = f(\omega)$ different non-linear transfer functions can be obtained with the same structure but with non-linear elements having different laws (but the same relations between one another). Alternative rectangle-diagram representations such as V_pI_q or even V_1I_1 (where V_1 and I_1 are projections on to arbitrary voltage and current reference axes) will give rise to non-linear circuits having the same structure but different elements. In general, the non-linear elements will have characteristics not easily realizable; the analogy might be useful in analysis.

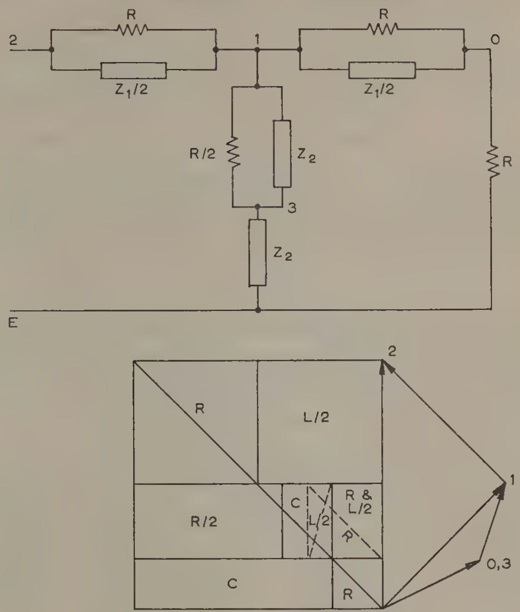


Fig. 3.—Constant-resistance T-section.

The complexity of the general non-linear network arises from the fact that the values of the non-linear elements depend on local voltages or currents and not on an independent parameter. The class of non-linear networks which are analogous to a.c. networks in the sense described above are composed of non-linear resistors whose relationships with one another are determined by a parameter ω. A different simplification results when a common current flows through all the elements or combinations of elements that comprise a network. Thus, analogous to the expansion of an impedance function into Foster elements, a non-linear resistance $A(i)/B(i)$ may be expanded in partial fractions to give a series of elements having resistances of the form $i/(i + \alpha)$. Alternative forms would be $1/(i + \alpha)$ or $(i + a)/(i + \alpha)$, where $(i + a)$ would appear in the numerator of each element. Such an expansion would be useful only if elements having such a resistance could be easily realized and the value of α simply altered.

621.391.812.7: 621.396.621

The Institution of Electrical Enginee
Paper No. 3221
Mar. 19

V.H.F. SOUND BROADCASTING

Subjective Appraisal of Distortion due to Multi-Path Propagation in F.M. Reception

By R. V. HARVEY, B.Sc., Associate Member.

(*The paper was received* 18*th February*, 1959. *It was published in March*, 1960, *and was read before the* ELECTRONICS AND COMMUNICATIONS SECTION 4*th April*, 1960.)

SUMMARY

In f.m. reception the delayed signals caused by multi-path propagation result in unwanted amplitude and phase modulation of the primary signal, and consequent distortion of the programme output of the receiver. The paper describes the results of tests which were carried out in simulated multi-path conditions to determine the importance of the parameters of both the received signal and the receiver in influencing the subjective annoyance caused by the distortion.

With a well-designed receiver, the distortion of piano music is 'slightly disturbing' when a single delayed signal is present having an equivalent path difference of 8 km and an amplitude of 35% relative to the primary signal. For a path difference of 29 km, however, the permissible relative amplitude is only 6% for the same subjective annoyance. Under the same conditions the distortion of speech is imperceptible. In comparison, receivers providing inadequate suppression of the unwanted amplitude modulation are much more susceptible to the distortion.

The use of pre- and de-emphasis appreciably reduces the distortion, being equivalent to a reduction of about 8 dB in the amplitude of the delayed signal when the path difference is about 16 km. Similarly, the distortion is less noticeable when the loudspeaker has a poor response at high audio frequencies.

The mechanism of multi-path distortion is discussed, and the harmonic spectra of the distortion shown for particular conditions.

(1) INTRODUCTION

It is well known that, when receiving an f.m. transmission, the presence of delayed signals results in distortion of the programme. As the transmitter frequency changes, the variation of the phase difference between the primary and delayed signals produces both unwanted amplitude and phase modulation of the primary signal. The phase-modulation component is physically indistinguishable from the wanted frequency modulation, and appears at the output of an ideal f.m. receiver as if it were part of the programme. On the other hand, the amplitude modulation, which can increase the amount of distortion if it is allowed to affect the receiver output, can be almost completely suppressed in a receiver of good design. Hayes and Page[1] have given a minimum value for the degree of suppression of amplitude modulation necessary to avoid this increase in distortion. The present paper is an extension of their work and is concerned with the audibility of the distortion for different ratios of the primary and delayed signals, and the manner in which the subjective annoyance changes with the receiver design parameters. The audible distortion caused by delayed signals varies from a noise similar to that of an overloaded a.f. amplifier when the path difference is small to a hiss like co-channel interference when the path difference is large. At intermediate path differences the distortion closely resembles the effect of something loose in the loudspeaker.

An extensive treatment of delayed-signal distortion has been given[2] and a comparison made between theory and objective measurements, but these results are not directly applicable to th conditions considered here. No attempt is made to produce theory to account for the results of subjective tests, but som waveforms and spectra are given in Section 8 to illustrate th mechanism of distortion of a single-tone modulation by a sing delayed signal in an imperfect receiver.

The subjective tests were carried out using a composite sign consisting of a primary signal together with a single delaye signal of controllable amplitude and delay. Provision for co trol of the relative phase at the carrier frequency was also mad but this did not affect the subjective distortion, apart from slight variation at the smallest delay used. For the majority o the tests the modulation was piano music, since this programm material had been found to be the most susceptible to distortio One of the tests was repeated with speech modulation, which relatively insensitive to this type of distortion.

The receiver parameters varied were the a.m. suppressio ratio, the type of a.m. suppression characteristic and the lou speaker response. The effects of pre- and de-emphasis and c changes in the degree of modulation were also investigated.

(2) EXPERIMENTAL PROCEDURE

(2.1) Generation of Primary Signal

A high-grade tape recording of a piano recital was used a the main test programme; the selected portion had little long term variation in level, so that comparison of two conditions i succession could be made with consistent results.

The programme was arranged to modulate a 91·3 Mc/s carrie from a signal generator either with pre-emphasis (50 microse time-constant), or without, as required. In tests on the effec of pre-emphasis, the switching-in of the network accentuatin the higher frequencies was accompanied by a 4 dB reduction i input level, in accordance with standard B.B.C. line-up practice

(2.2) Simulation of Multi-Path Propagation Conditions

Multi-path propagation conditions were simulated by sendin part of the primary signal into a terminated delay line and extract ing a signal with the desired delay by means of a probe. The delayed signal was then combined with the primary signal ir the desired proportion and the composite signal applied to the receiver input.

The delay line was a 30 m length of special cable of Germar manufacture having a delay of 3·3 microsec/m; its inner conductor consisted of a helix of fine wire on a flexible magnetic core, and its outer conductor, separated from the inner by a layer of polystyrene tape, consisted of a layer of longitudina conductors forming an electrostatic screen. The pick-up probes were coils of wire wound over the screen to couple with the magnetic field of the inner conductor; these were at positions giving delays corresponding to path differences of 3·2 km, 8 km, 16 km and 29 km. Normally the cable was terminated by a resistor equal to its characteristic impedance of 4 kilohms, but, by short-circuiting the termination and picking up from a direc-

Mr. Harvey is with the British Broadcasting Corporation.

nal coupler near the sending end, a delay corresponding to km could be obtained. The cable was mounted in a large x and cross-threaded in three dimensions to minimize wanted magnetic coupling between adjacent parts.

As the attenuation of the delay line was too large to permit use directly at 91·3 Mc/s, the frequency of the primary signal as first converted to 1 Mc/s before delay, and the composite gnal converted back to 91·3 Mc/s; the double conversion was fected by one local oscillator and two balanced crystal modu-tors as shown in the schematic of Fig. 1.

(2.3) Control of Parameters of System

The parameters of the receiving and sound reproducing system which may affect the subjective distortion are discussed with particular reference to the two receivers, A and B, employed in the tests.

(2.3.1) A.M. Suppression Ratio.

The efficiency of limiting in the receiver can be measured as the ratio between the a.f. output due to an f.m. signal and that due to an a.m. signal. This suppression ratio was measured

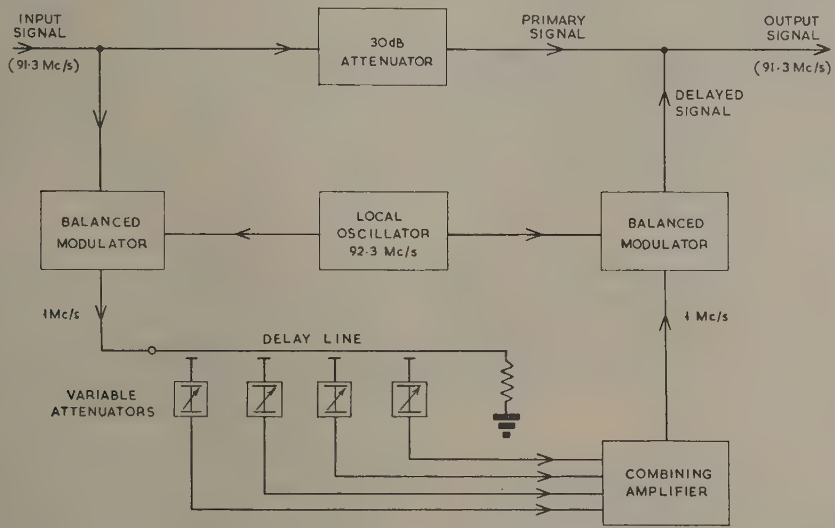


Fig. 1.—Schematic of multi-path propagation simulator.

In a later modification of the apparatus, the signal delays were also obtained at 1 Mc/s, but, instead of a cable, use was made of a 45 cm length of nickel wire carrying longitudinal waves excited by magnetostriction. The wave was launched and picked up by small coils mounted on the wire, with a small permanent magnet near each coil to give longitudinal polarization. In these tests the maximum effective path difference was 24 km.

In order to measure the relative amplitude and delay of the signals produced by the simulator, the composite signal was also fed, in parallel with the receiver used for the listening tests, to a special receiver designed for multi-path propagation investigation.

This receiver contains a cathode-ray-tube display of the instantaneous amplitude of the input signal as a function of its instantaneous frequency. The vertical deflection is derived by a d.c. coupled amplifier from the amplitude-limiting circuit and the horizontal deflection is derived from the discriminator output. In the absence of delayed signals, a frequency-modulated signal of constant amplitude produces a horizontal straight line of a length proportional to its peak deviation. If one delayed signal is present, the line becomes an approximate sinusoid, as shown in Fig. 2. Here, the deviation is ± 75 kc/s at 120 c/s; the base line is traced by removing the vertical deflection voltage and corresponds to zero signal amplitude. It can be shown that, if n complete cycles are displayed by the trace in a total deviation of ± 75 kc/s, the equivalent path difference of the delayed signal is $2n$ kilometres. The display of Fig. 2 is therefore readily interpreted as indicating a delayed signal having an equivalent path difference of 8 km and a relative amplitude, a/b, of 20% of that of the primary signal.

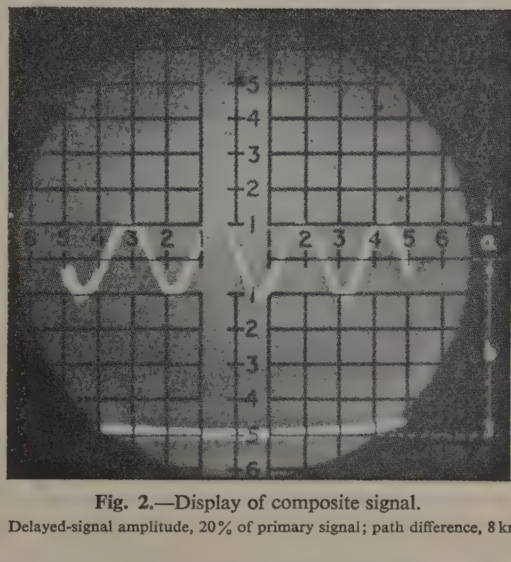


Fig. 2.—Display of composite signal.
Delayed-signal amplitude, 20% of primary signal; path difference, 8 km.

with 40% amplitude modulation and ± 30 kc/s frequency modulation (i.e. 40% of the system deviation) applied simultaneously, using different audio frequencies in order to separate the a.m. and f.m. outputs. This method provides a single ratio which has been found to have a reasonable correlation with the effectiveness of a.m. suppression for most practical purposes.

Receiver A was specially adapted to enable the a.m. suppres-

sion ratio to be varied from 20 to 50 dB by altering the pre-limiter gain. Receiver B was a current domestic model in the popular price range; it had an a.m. suppression ratio of 16 dB but, by switching in a specially fitted limiter,[3] the a.m. suppression ratio could be improved to 26 dB.

(2.3.2) **Type of A.M. Suppression Characteristic.**

Most receivers exhibit maximum a.m. suppression near the centre of the pass band, with the a.m. response increasing on either side. Ideally the receiver is tuned so that maximum suppression occurs at the carrier frequency, but, in practice, other considerations may cause some mistuning.

Receiver A was capable of operating in two conditions, the first showing an a.m. response which was zero at the carrier frequency but increased linearly on either side, and the second showing a constant a.m. suppression over the pass band. The two conditions were such as to give the same suppression ratio (20 dB) as defined in Section 2.3.1. Receiver B incorporated a ratio detector and had the normal type of suppression characteristic, with the a.m. response increasing on either side of the tuning point.

The ratio detector is particularly susceptible to a special case of multi-path propagation, namely that due to reflection from an aeroplane in flight, usually referred to as aircraft flutter. This can cause an annoying form of distortion if, for the particular strength and rate of flutter of the received signal, the design of receiver is such that a.g.c. or pre-detector limiting does not operate. This aspect of a.m. suppression was not included in the investigation described in the paper.

(2.3.3) **Effect of Pre-emphasis and De-emphasis.**

The B.B.C. uses a system incorporating pre-emphasis of the higher modulating frequencies by a network having a time-constant of 50 microsec. In order to preserve the peak deviation of ± 75 kc/s, the transmitter deviation is 4 dB lower at low modulating frequencies and 6 dB higher at 10 kc/s than that of a system without pre-emphasis but with the same peak deviation.

Since much of the audible distortion produced by delayed signals consists of high harmonics of low modulating frequencies, both the reduced deviation at low frequencies and the attenuation of high frequencies by the de-emphasis in the receiver might be expected to have an important effect on the annoyance caused. Tests were therefore made to assess the advantage gained by the use of pre-emphasis.

(2.3.4) **Type of Loudspeaker.**

All the listening tests with receiver A were made with a high-quality monitoring loudspeaker having a good frequency response from 50 c/s to 12 kc/s. Receiver B was tested both with its own loudspeaker, a unit typical of that used in current domestic models, and also with the high-quality unit, in order to assess the effect of the improved high-frequency response.

(2.4) Arrangements for Subjective Tests

The tests were of two types, namely comparative tests, during which groups of several subjects were asked to compare the distortion in one condition with that in a reference condition, and absolute tests, in which engineers were asked to establish conditions which they assessed as conforming to each of four subjective grades of distortion.

(2.4.1) **Comparative Tests.**

A group of about six subjects was arranged before the loudspeaker and presented with a continuous recital of piano music, during which the system alternated between a test condition and the reference condition, identified by coloured lights.

Preliminary tests established, at each value of path differen two values of the delayed-signal amplitude, giving in one ca more and in the other less annoyance than the reference co dition, without being so nearly equivalent as to tax the judgme of the subject too severely. The change over between conditio was made automatically every 4 sec, this interval proving abo optimum for ease of comparison. The subjects were asked note the more annoying condition, the choice being indicat by pressing a 3-position key to the appropriate colour, or leavi the key in the neutral position if the two conditions appear equally bad. The total votes cast in each pair of compariso were used as weighting factors to estimate the test conditi which would have sounded equivalent to the reference conditio

In this way, one 10 min series of ten comparisons gave t average opinion of six subjects on five combinations of t delayed-signal amplitude and path difference which wou result in equal annoyance. These combinations can be plott on a graph to give a curve of equal annoyance; such a cur will be referred to as an isolype.*

(2.4.2) **Absolute Tests.**

To assess the degree of annoyance caused by different types distortion, individual engineers were asked to judge each co dition in terms of the four subjective criteria: just perceptibl perceptible, slightly disturbing and disturbing.

In this test, an uncalibrated knob was provided for controlli the amplitude of the delayed signal; the distortion cou thus be varied from imperceptible to intolerable. The kn was adjusted for each criterion in turn, the undistorted pr gramme being available for comparison by the operation of switch; the completion of the adjustment was indicated b pressing a button. The operator recorded the settings chose and then presented the subject with a different test condition.

This procedure was adopted only for engineers since it require some skill in setting a knob to produce a desired change. Give this skill, the subject was less distracted than in the comparati tests and the results were found to be very consistent. similar procedure can also be used by individual engineers f equating a test condition with a reference condition.

(3) RESULTS OF TESTS

(3.1) Performance of a Good Receiver

Fig. 3(*a*) shows a comparative isolype for receiver A; th a.m. suppression ratio was 50 dB and a high-quality loudspeake was used. The reference condition is represented by the poi (16 km, 10%), and the other points represent the mean opinio of 42 listeners as to the combinations of amplitude and pat difference giving annoyance equal to that for the referenc condition.

Fig. 3(*b*) shows absolute isolypes for the four subjecti criteria, the listeners being six experienced engineers. The com parative isolype of Fig. 3(*a*) lies between the 'slightly disturbin and 'disturbing' curves in Fig. 3(*b*).

Fig. 3(*c*) shows the corresponding results using a programm consisting of speech instead of piano music. For equall annoying distortion of speech, the delayed signal amplitude i higher than for piano music, the ratio being about 11 dB a path differences near 16 km.

(3.2) Performance of a Receiver with Imperfect Limiting

Figs. 4(*a*) and (*b*) show corresponding isolypes for receiver B which had an a.m. suppression ratio of 16 dB and incorporate a small loudspeaker. The test procedure was the same as tha used for receiver A but the reference condition was changed t

* From the Greek *isos* (equal) and *lype* (pain, discomfort).

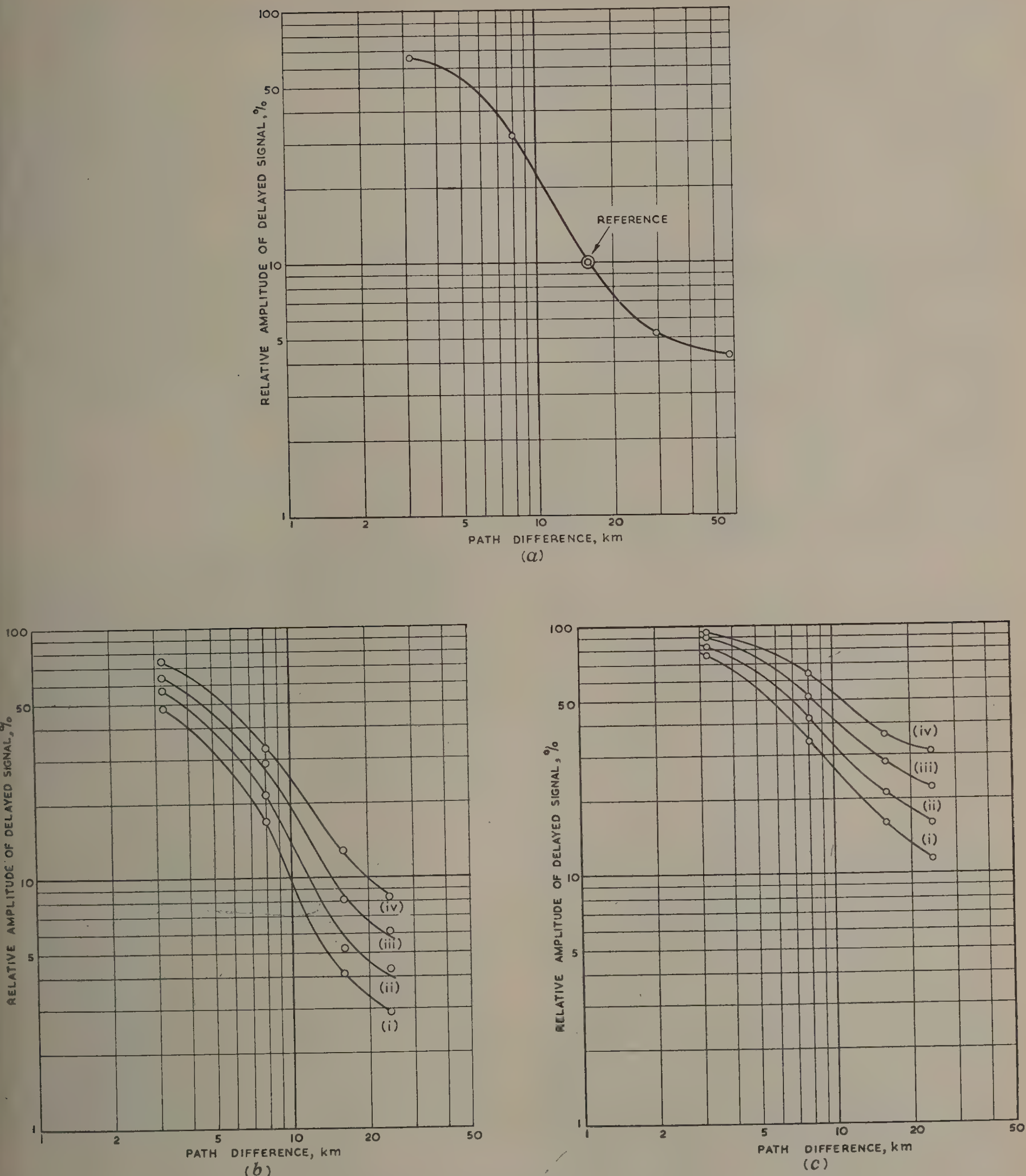
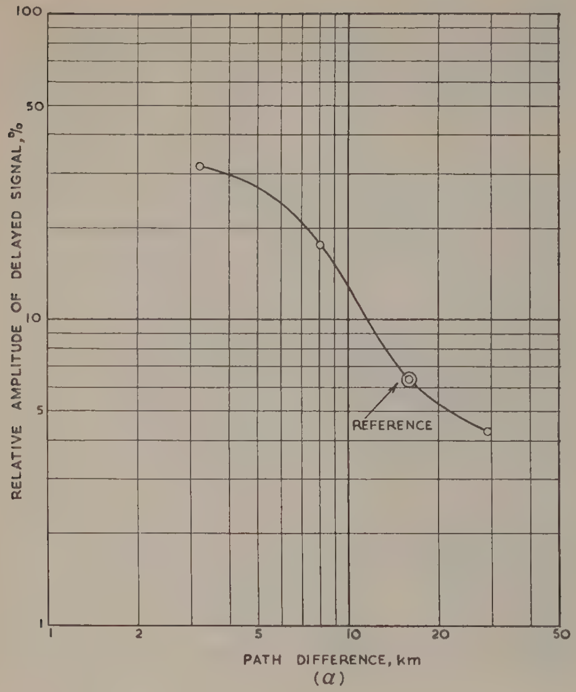


Fig. 3.—Isolypes for receiver A.

(a) Comparative isolype for piano-music modulation.
(b) Absolute isolypes for piano-music modulation.
(c) Absolute isolypes for speech modulation.

(i) Just perceptible. (ii) Perceptible.
(iii) Slightly disturbing. (iv) Disturbing.

A.M. suppression ratio: 50 dB.
Loudspeaker: high-quality unit.



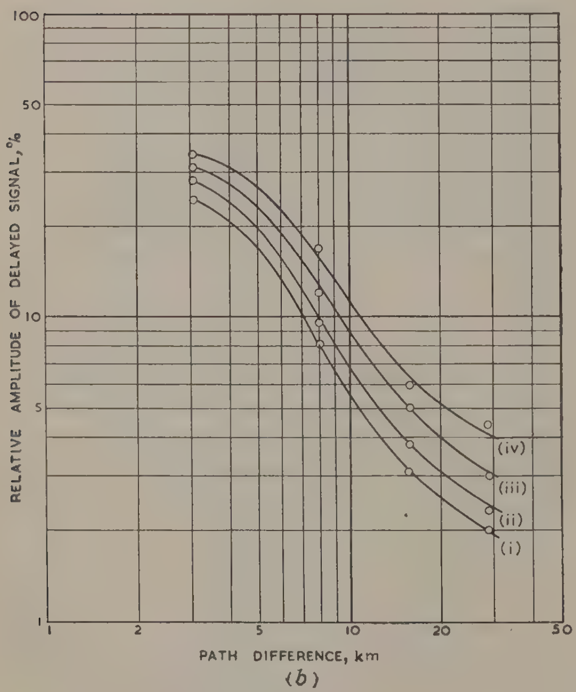


Fig. 4.—Isolypes for receiver B.

(*a*) Comparative. (*b*) Absolute.

(i) Just perceptible.
(ii) Perceptible.
(iii) Slightly disturbing.
(iv) Disturbing.

A.M. suppression ratio: 16 dB.
Modulation: piano music.
Loudspeaker: 10 in elliptical.

(16 km, 6·3%) for the comparative curve of Fig. 4(*a*); th[illegible] nearly coincides with the absolute isolype for the criteri[illegible] 'disturbing' in Fig. 4(*b*).

(3.3) Comparison of Receivers A and B using the Same Loudspeaker

The receivers used in the tests described in Sections 3.1 ar[illegible] 3.2 were compared when connected alternately to the san[illegible] high-quality loudspeaker. A slight low-frequency loss w[illegible] deliberately introduced in receiver A to eliminate differences [illegible] the quality of reproduction of the two receivers in the absen[illegible] of delayed signals. The object of the tests was to find, f[illegible] receiver B, an isolype equivalent to that found for receiver A[illegible] given in Fig. 3(*a*). Fig. 5 shows the result of a comparis[illegible]

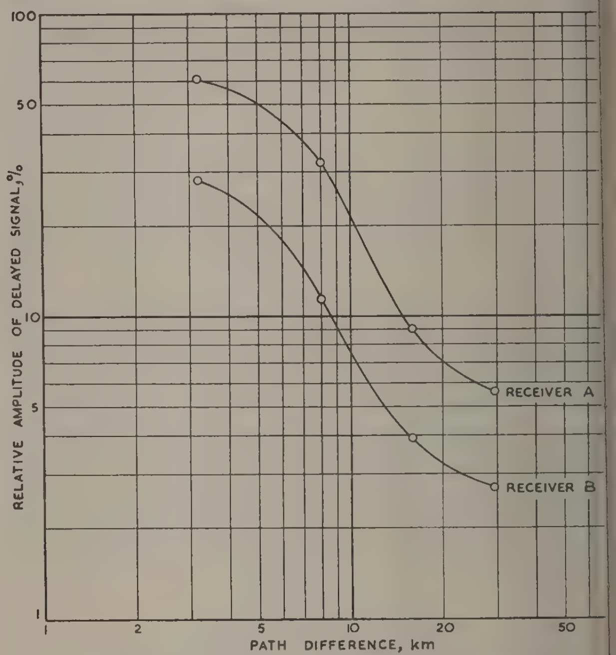


Fig. 5.—Direct comparison of receivers A and B.

A.M. suppression ratio: Receiver A, 50 dB.
Receiver B, 16 dB.
Modulation: piano music.
Loudspeaker: high-quality unit for both receivers.

made by several engineers. For each selected path difference the lower curve gives the amplitude of delayed signal required for the output of receiver B to sound as distorted as that of receiver A, the delayed-signal amplitude for receiver A having been set to a value following the isolype of Fig. 3(*a*).

Although the signals being compared had the same path difference in each case, the type of distortion was not quite the same, since that caused by amplitude modulation, present only in the output of receiver B, had a different spectrum from that of the phase distortion.

(3.4) Effect of increasing the A.M. Suppression Ratio

Fig. 6 shows isolypes relating the delayed-signal amplitude to the a.m. suppression ratio of the receiver at each of four values of path difference, the subjective distortion being 'slightly disturbing'. Results were obtained by comparing different conditions of receiver A having a.m. suppression ratios of 20, 30 and 50 dB, the change in suppression being effected by a change

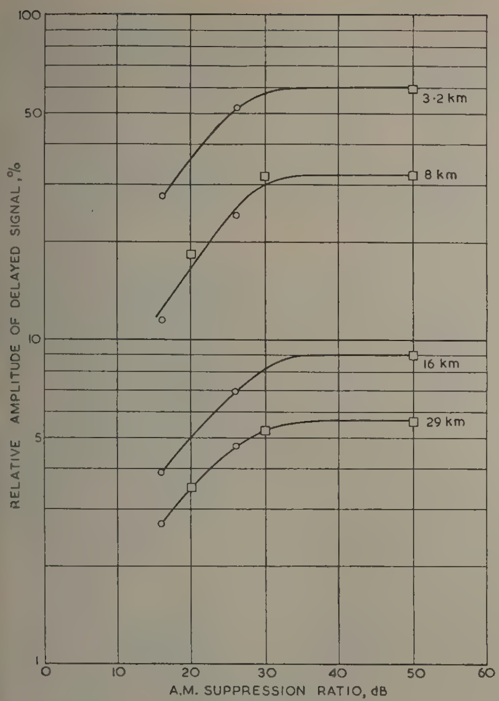


Fig. 6.—Isolypes showing effect of increasing a.m. suppression ratio.

Modulation: piano music.
Loudspeaker: high-quality unit for both receivers.
□ Receiver A. ○ Receiver B.

of pre-limiter gain. The results of the tests in Section 3.3 using receiver B are also shown; in addition, a further test was made in which receiver B was operated with an extra diode limiter which increased the a.m. suppression from 16 to 26 dB.

The curves indicate a considerable improvement in performance as the a.m. suppression ratio is increased up to 30 dB, but little improvement at higher suppression ratios.

(3.5) Effect of Changing Type of A.M. Suppression Characteristic

Receiver A was adapted to have two alternative a.m. suppression characteristics, although having the same a.m. suppression ratio as measured by the method described in Section 2.3.1. The first characteristic was such that the amplitude modulation appearing at the output was zero at the mean carrier frequency and increased linearly with deviation from this frequency. This type of a.m. response characterizes all balanced discriminator circuits, such as the Foster–Seeley circuit used in receiver A. The second type of characteristic was such that the a.m. response was independent of deviation. This condition may be approached in a mistuned or badly aligned receiver; it was simulated by first adjusting receiver A to have an a.m. suppression ratio of 50 dB and then adding to the output a small a.m. signal from the grid circuit of the limiter valve.

Arrangements were made to switch the receiver from one condition to the other, both conditions having been adjusted to have an a.m. suppression ratio of 20 dB. The results of a comparison by six engineers are given in Table 1.

The Table shows that, of the two receivers with the same nominal suppression ratio, the performance of the receiver with a uniform a.m. response was about 4 dB poorer than that of

Table 1

Path difference	Delayed-signal amplitude for same annoyance		Ratio
	A.M. response proportional to deviation	A.M. response uniform	
km	%	%	dB
3·2	57	46	1·8
8	25	16	4·0
16	9	5·5	4·2
29	5	3·1	4·2

the receiver with a balanced discriminator, for delayed signals with a path difference of 8 km or more. The most noticeable effect of the uniform a.m. response was the greater a.m. distortion at low modulation levels and large path differences.

(3.6) Effect of Removing Pre-emphasis and De-emphasis

The pre-emphasis network in the modulating circuit could be interchanged with an attenuator in the output circuit of the receiver. By this means, the modulation level at low frequencies was increased by the requisite 4 dB, the pre-emphasis was removed and the de-emphasis in the receiver cancelled, while preserving the overall gain of the system.

A comparison was made of the delayed-signal amplitudes which gave equal annoyance in the two conditions using receiver A, operating with a 50 dB a.m. suppression ratio; the results are given in Table 2.

Table 2

Path difference	Delayed-signal amplitude for same annoyance		Ratio
	F.M. system with 50 microsec pre-emphasis	F.M. system with no pre-emphasis	
km	%	%	dB
8	28	11	8
16	9	3·6	8
29	5	2·2	7

It is seen that the use of 50 microsec pre-emphasis (with de-emphasis at the receiver) enables a 7–8 dB larger amplitude of delayed signals to be tolerated when the path difference is in the 8–29 km range. The most noticeable difference in character is the greater preponderance of high-frequency distortion in the absence of pre-emphasis and de-emphasis.

(3.7) Effect of Over-modulation and Under-modulation

If the transmitter deviation is doubled and the receiver output halved to preserve the same programme level, the distortion caused by a small delayed signal is expected from theory to be the same as that caused at the original transmitter deviation by a signal of twice the delay and half the amplitude (see Section 8). This result was confirmed by tests on receiver A whose bandwidth had been made great enough to accept ± 150 kc/s deviation with negligible distortion. Similarly, under-modulation produced the inverse effect. The above equivalence breaks down for delayed signals of large relative amplitude or long delay, but appears to be adequate for most practical conditions.

(4) DISCUSSION OF RESULTS

A feature of the isolypes of Figs. 3 and 4 is the extent to which delayed signals become rapidly more annoying as the path

difference increases. For path differences less than about 2 km, which are below the range investigated, delayed signals of relatively large amplitude can be tolerated so long as adequate a.m. suppression is maintained. However, in the case of the ratio detector, a delayed signal greater than about 50% of the primary signal will often cause distortion unless extra limiting is provided in front of the detector. This is a result of a diode cut-off effect which can occur when the signal applied to the ratio detector has a high percentage of amplitude modulation.

As the path difference increases from 2 km, the spectrum of the distortion extends to higher-order harmonics; this is clear from the example given in Section 8. It would appear that the ear is very sensitive to high-order harmonics, since the tolerable amplitude of delayed signal falls very rapidly over the range of path difference from 8 to 20 km. For path differences of 30 km and above, the harmonic spectrum extends beyond the audible range and the distortion is heard mainly as a hissing noise. Very large path differences (100 km or more) are met only in exceptional circumstances, for example by reflection from intense aurora. The conditions are then practically identical with those of interference from a co-channel station carrying the same programme. In this case, using a receiver with good a.m. suppression, experiments indicate that the distortion of piano music is perceptible in the presence of an unwanted signal equal to 1·2% of the main signal. This result is to be compared with about 4% for a path difference of 29 km as shown in Fig. 3(*b*). The decrease in the tolerable level of the delayed signal evidently continues as the path difference increases, and presumably approaches the co-channel interference figure above 100 km.

Another important feature of Figs. 3 and 4 is the way in which the annoyance varies with the amplitude of the delayed signal; a doubling of the amplitude, for instance, changes the annoyance by two subjective grades when the path difference is 10 km.

Most of the work described has been carried out with a programme of piano music because of its high susceptibility to this type of distortion; the difference between piano music and other types of programme is, in fact, quite striking. Thus, Figs. 3(*b*) and (*c*) show that, under certain conditions, distortion can be 'just perceptible' on speech yet worse than 'disturbing' on piano music. In general, for the same grade of distortion of speech and piano music, the ratio of the delayed-signal amplitudes increases with the path difference and is about 11 dB for a path difference of 16 km. A similar effect was observed with co-channel interference, the corresponding ratio in this case being 16 dB.

Turning now to the importance of a.m. suppression in the receiver, Fig. 6 shows that a receiver giving an a.m. suppression ratio of 30 dB is practically as good as one with a very high suppression ratio. A reduction to 16 dB, on the other hand, is equivalent to doubling the amplitude of the delayed signal, or a deterioration of two grades in the subjective assessment.

In these tests, the suppression ratio is measured by a method involving simultaneous amplitude and frequency modulation, as this method gives a single figure for the suppression which is reasonably consistent with the subjective assessment. For example, two receivers having extremely different types of a.m. suppression characteristic, but the same measured a.m. suppression ratio, differed in their sensitivity to distortion by only 4 dB.

It is preferable for a receiver to give an a.m. suppression ratio of at least 35 dB by the simultaneous modulation method, when carefully tuned according to the manufacturer's instructions, in order to be reasonably sure of optimum performance under multi-path conditions. This makes a small allowance for tuning inaccuracy or oscillator drift, which will generally cause some deterioration in a.m. suppression. An alternative method of measuring the a.m. suppression ratio which is sometimes used is to compare the response of the receiver in turn to equal percentages of frequency and amplitude modulation (usually 30% modulation). If this sequential method is used, a very high suppression ratio at the carrier frequency to which the receiver is tuned does not alone ensure a minimum of multi-path distortion. To ensure reasonable protection, a figure of 30 dB or more should be obtained not only at the carrier frequency but also when the test is applied at frequencies ±30 kc/s relative to the carrier frequency. (A receiver without automatic frequency control is assumed.)

The advantage conferred by the use of pre- and de-emphasis under multi-path conditions is interesting. The delayed-signal amplitude can be increased by some 8 dB (Table 2) before it causes the same degree of annoyance as in a system without pre-emphasis; on the other hand, the improvement in signal/hiss ratio due to pre-emphasis in the absence of multi-path effects is only about 2 dB.* The reason for this may be appreciated if we refer to the effect of over- or under-modulation mentioned in Section 3.7. When pre-emphasis is applied, the deviation at low audio frequencies is reduced by 4 dB. We may therefore consider its effect in two steps; (*a*) a reduction of deviation at all audio frequencies by 4 dB and (*b*) an increase of deviation at high audio frequencies according to the standard 50 microsec pre-emphasis, together with the introduction of the corresponding de-emphasis in the receiver. Step (*a*) is equivalent to an increase in amplitude of the delayed signal in the ratio of 1·6 : 1 and a reduction in path difference in the same ratio. Fig. 3(*a*) shows that the results of these two changes, starting at the reference point (16 km, 10%), is to give a new point (10 km, 16%) below the curve, corresponding to a 2·7 dB improvement. Step (*b*), the application of pre- and de-emphasis without change of deviation at low frequencies, is expected to produce an improvement comparable with that for wide-band noise, namely 6 dB. Thus a total increase of 8–9 dB in the delayed-signal amplitude should be permissible when the path difference is about 16 km. Step (*a*) in the case of receiver hiss, of course, gives a 4 dB loss of signal/hiss ratio so that the net improvement is only 2 dB.

The example given above has assumed a condition where a decrease in path difference is most potent in reducing distortion. For path differences outside the range covered by Table 2, the effect of reduced deviation may no longer be beneficial and the improvement due to pre-emphasis is expected to fall to 6 dB or even less.

A decrease in the loudspeaker response at the higher audio frequencies also reduces the audible distortion; the reduction in changing from the high-quality loudspeaker to the 10 in elliptical unit in receiver B was equivalent to a 2 dB reduction in the amplitude of the delayed signal.

(5) CONCLUSIONS

Subjective tests on the amount of distortion produced in an f.m. receiver when, in addition to the primary signal, a delayed signal is present, have led to the following main conclusions:

(*a*) A marked rise in sensitivity to distortion occurs as the delay is increased, especially for delays corresponding to a range of path differences of 8–16 km.

(*b*) A delayed-signal amplitude of as little as 10% of the primary signal can cause 'perceptible' distortion of piano music with a good receiver for a path difference of 12 km; moreover, the distortion under the same conditions may exceed the 'disturbing' level in some receivers if the a.m. suppression is insufficient. It is recommended that the receiver shall be designed to

* This figure is based on subjective measurements using a wide-range loudspeaker and takes into account the 4 dB reduction of deviation at low frequencies which is mentioned in Section 2.3.3.

give a suppression ratio of at least 35 dB (measured by the simultaneous method) to ensure a performance close to that of an ideal f.m. receiver.

(*c*) For speech and many other types of programme, much higher delayed-signal amplitudes can be tolerated than for piano music.

(*d*) A pre-emphasized system with a 50 microsec time-constant confers 7–8 dB greater protection against multi-path distortion than an f.m. system of similar peak deviation ($\pm$75 kc/s) without pre-emphasis.

The severity of distortion is reduced if the a.f. amplifier or loudspeaker in use has a poor treble response, and distortion may even escape detection under conditions when it would be quite obvious with a wide-range amplifier and loudspeaker.

(6) ACKNOWLEDGMENTS

The author is indebted to the Chief Engineer of the British Broadcasting Corporation for permission to publish the paper, to Dr. G. J. Phillips for many helpful discussions, to Mr. J. G. Spencer for providing the multi-path display receiver mentioned in Section 2.2 and to the staff of the B.B.C. Research Department for their patience in submitting to subjective tests.

(7) REFERENCES

(1) Hayes, E. W., and Page, H.: 'The B.B.C. Sound Broadcasting Service on Very High Frequencies', *Proceedings I.E.E.*, Paper No. 2258 R, November, 1956 (**104** B, p. 213).

(2) Medhurst, R. G., and Small, G. F.: 'An Extended Analysis of Echo Distortion in the F.M. Transmission of Frequency-Division Multiplex', *ibid.*, Paper No. 2006 R, March, 1956 (**103** B, p. 190).

(3) Mural, F.: 'Dynamic-Diode Limiter for F.M. Demodulators', *Electronics*, August, 1955, **28**, p. 146.

(4) Lewin, L., Muller, J. J., and Basard, R.: 'Phase Distortion in Feeders', *Wireless Engineer*, 1950, **27**, p. 143.

(8) APPENDICES

(8.1) Distortion Caused by a Delayed Signal

The effect of a delayed signal in producing the unwanted amplitude and phase modulation of the composite signal is illustrated in Fig. 7. It is supposed that the primary signal is

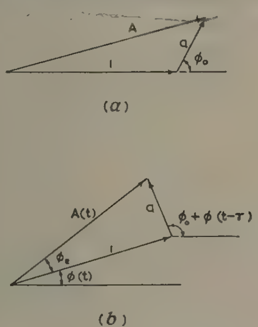


Fig. 7.—Vector diagram showing effect of delayed signal.
(*a*) Unmodulated carrier.
(*b*) Frequency-modulated carrier.

of unit amplitude and that a second signal of amplitude a, travelling by a longer path, is added to it at the receiver. In the absence of modulation, the relative phase of the second signal will take a value ϕ_0, which depends upon the amount by which the path difference exceeds an integral number of wavelengths at the carrier frequency, and includes any phase change occurring on reflection. The received signal thus differs from the primary signal in both phase and amplitude as shown by the resultant, A, in Fig. 7(*a*).

When frequency modulation is applied, it is convenient to represent the modulation of the received primary signal by the phase deviation $\phi(t)$. Fig. 7(*b*) shows the vector diagram for an instant, t. The delayed signal has changed in phase by $\phi(t-\tau)$ because it traverses the longer path and is delayed by a time τ relative to the primary signal. Since the primary signal is distortionless, we may regard the angle ϕ_e as the phase error, or phase distortion, in the composite signal. Making the quasi-stationary approximation for the case where $a \ll 1$, we may write for the resultant:

$$A(t) = 1 + a\cos[\phi_0 + \phi(t-\tau) - \phi(t)] \quad . \quad . \quad . \quad . \quad (1)$$

$$\phi_e(t) = a\sin[\phi_0 + \phi(t-\tau) - \phi(t)] \quad . \quad . \quad . \quad . \quad . \quad (2)$$

$$\omega_e(t) = \frac{d}{dt}\phi_e(t) = a\cos[\phi_0 + \phi(t-\tau) - \phi(t)] \times [\omega(t-\tau) - \omega(t)] \quad (3)$$

Eqn. (3) gives the distortion term $\omega_e(t)$ of the instantaneous angular frequency defined by $\omega = d\phi/dt$; the distortion appearing at the output of a discriminator will be proportional to this term, although it will afterwards be modified if de-emphasis is applied. Insufficient a.m. suppression will allow a further contribution to the output, but this will follow eqn. (1) closely in form only if the a.m. suppression is constant over the deviation range.

(8.2) Distortion Waveforms and Spectra for Sinusoidal Modulation

An analysis of the harmonic distortion produced by a single delayed signal has been given for a system with perfect amplitude limiting.[4] A similar analysis is developed here but is extended to include the effects of imperfect limiting in certain types of receiver.

The above equations are quite general with regard to the modulation applied; however, it is interesting to examine the result if a simple sine-wave modulation is applied. Thus, referring to eqn. (1), we put

$$\omega(t) = \omega_d \cos pt$$

or

$$\phi(t) = \phi_d \sin pt$$

where $\phi_d = \omega_d/p$, and insert the values $a = 0{\cdot}1$, $p = 10^3\pi$, $\phi_0 = \pi/4$, $\phi_d = 60$, $\tau = 67$ microsec and $\omega_d = 6 \times 10^4\pi$, i.e. $\pm$ 30 kc/s peak deviation at 500 c/s is applied, the delayed-signal amplitude is 10% of the primary signal and the path difference is approximately 20 km, but actually is such as to correspond to the chosen carrier phase difference of $\pi/4$.

Fig. 8 shows the waveforms for this case; (*a*) shows the instantaneous frequency as a function of time for the primary and delayed signals, (*b*) the variation of amplitude with time given by eqn. (1) and (*c*) the same variation of amplitude plotted against the instantaneous frequency of the primary signal, $\omega(t)$.

For sinusoidal modulation, eqn. (1) becomes

$$A(t) - 1 = a\cos[\phi_0 - 2\phi_d \sin\tfrac{1}{2}p\tau \cos p(t - \tfrac{1}{2}\tau)]$$

$$= a\cos\left[\phi_0 - \frac{2}{p}(\sin\tfrac{1}{2}p\tau)\omega(t - \tfrac{1}{2}\tau)\right] \quad . \quad . \quad (4)$$

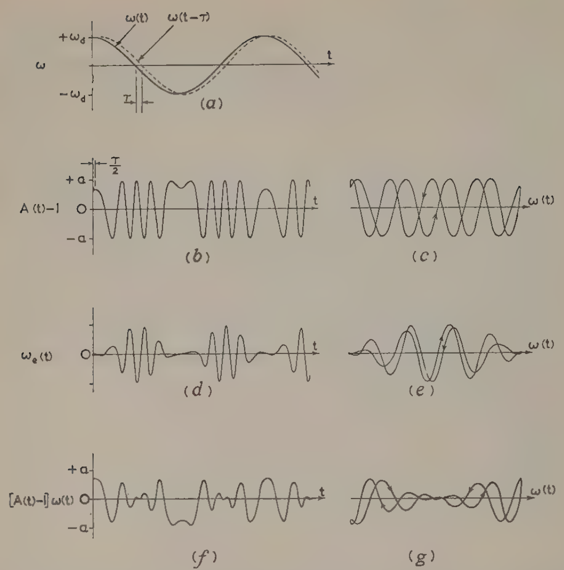


Fig. 8.—Distortion as function of time and frequency for sinusoidal modulation.

(*a*) Frequency of primary and delayed signals.
(*b*) Variation of a.m. component with time: uniform a.m. response.
(*c*) Variation of a.m. component with frequency: uniform a.m. response.
(*d*) Variation of f.m. component with time.
(*e*) Variation of f.m. component with frequency.
(*f*) Variation of a.m. component with time: Foster–Seeley a.m. response.
(*g*) Variation of a.m. component with frequency: Foster–Seeley a.m. response.

The points of symmetry of the time function (*b*) are therefore delayed by a time $\frac{1}{2}\tau$ and the amplitude modulation is a single-valued function of $\omega(t - \frac{1}{2}\tau)$ rather than $\omega(t)$. If the amplitude modulation is plotted against $\omega(t)$, as in Fig. 8(*c*), the result will approach a single-valued function of the modulating frequency, p, only when $\omega(t)$ is almost equal to $\omega(t - \frac{1}{2}\tau)$; in general there will be a phase shift. Although in the present case this is only a fraction of a modulation cycle (6°), it appears as a large phase shift of the a.m. function shown in Fig. 8(*b*). The f.m. receiver described in Section 2.2 enables the path difference of the principal delayed signals to be estimated. With a low modulation frequency a clear display is obtained, as shown in Fig. 2, but on a typical broadcast programme, the mis-phasing effect of the kind illustrated in Fig. 8(*c*) causes some blurring of the display.

Figs. 8(*d*) and (*e*) show the f.m. distortion term of eqn. (3) plotted in the same way. The term $\omega(t - \tau) - \omega(t)$ in the present example becomes

$$\omega_d\,[\cos p(t - \tau) - \cos pt] = 2\omega_d \sin \tfrac{1}{2}p\tau \sin p\,(t - \tfrac{1}{2}\tau) \quad (5)$$

From eqns. (3) and (5) the peak amplitude of the distortion is therefore $2a\omega_d \sin \frac{1}{2}p\tau$, which is $2{\cdot}1\%$ of the amplitude ω_d of the fundamental frequency modulation; the average percentage distortion would be about $1{\cdot}5\%$ owing to the sinusoidal modulation of the envelope of the distortion. Because of the high harmonic content this distortion figure is reduced somewhat by de-emphasis; it is still important, however, since this type of distortion is much more noticeable than low-order harmonic distortion, and it would have to be well under 1% to be inaudible.

Finally, Figs. 8(*f*) and (*g*) show the a.m. waveform as modified by an ideal Foster–Seeley discriminator. Here, as in the case of the unmodified amplitude modulation of (*b*) and (*c*), its importance in relation to the f.m. distortion is a function of

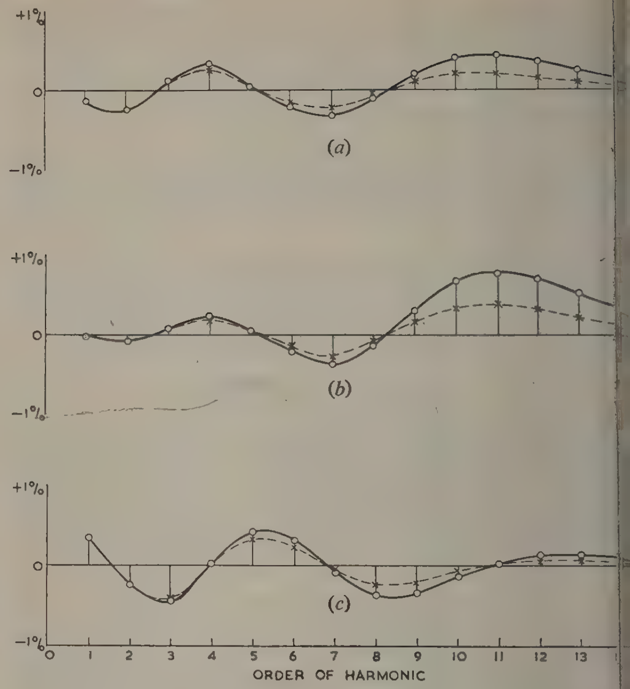


Fig. 9.—Distortion spectra for sinusoidal modulation.

(*a*) A.M. component: uniform a.m. response: a.m. suppression ratio, 20 dB.
(*b*) F.M. component.
(*c*) A.M. component: Foster–Seeley a.m. response: a.m. suppression rati 20 dB.

—○—○— No de-emphasis.
---×---×— 50 microsec de-emphasis.

the a.m. suppression; comparison is also complicated by th different spectral content. The spectra may be calculated b expanding expressions (4) and (5) in terms of Bessel coefficient and the result for the present example is given in Fig. 9.

The spectra shown in Figs. 9(*a*), (*b*) and (*c*) are those of th time functions (*b*), (*d*) and (*f*) of Fig. 8, but the a.m. spectra Figs. 9(*a*) and (*c*), have been reduced to correspond with an a.m suppression ratio of 20 dB, showing a power content comparabl with the f.m. spectrum in Fig. 9(*b*). The spectral envelope i shown as a full line for no de-emphasis, while the dashed lin shows how the spectrum is modified after 50 microse de-emphasis. The envelope is a smooth curve for the valu of ϕ_0 chosen in the example; for other values, alternat harmonics tend to predominate, but the redistribution of powe will have little effect on the subjective distortion.

(8.3) Effect of Change of Frequency Deviation

It is interesting to consider the effect on the distortion of change in the peak deviation of the primary signal. If th deviation is increased by a factor m, the phase deviation become

$$\phi'(t) = m\phi(t)$$

the a.m. distortion given in eqn. (1) becomes

$$A'(t) - 1 = a \cos\,[\phi_0 + m\phi(t - \tau) - m\phi(t)] \quad . \quad (6$$

and the phase distortion given by eqn. (2) becomes

$$\phi_e'(t) = a \sin\,[\phi_0 + m\phi(t - \tau) - m\phi(t)]$$

If, instead of increasing the deviation by m, the delay time, τ, increased and the delayed signal amplitude decreased by m, e same equations become

$$\phi''(t) = \phi(t)$$

$$A''(t) - 1 = \frac{a}{m}\cos\left[\phi_0 + \phi(t - m\tau) - \phi(t)\right] \quad . \quad (7)$$

$$\phi''_e(t) = \frac{a}{m}\sin\left[\phi_0 + \phi(t - m\tau) - \phi(t)\right]$$

omparing eqns. (6) and (7), the magnitudes of the distortion terms are equivalent in relation to the total deviation, subject to the original restriction $a \ll 1$.

The terms within brackets, which govern the form of the distortion, will be nearly equivalent so long as the delay time, τ, is short compared with one period of the highest significant modulating frequency. In this case, the rate of change of $\phi(t)$ will be almost constant during a time interval, τ, so that the terms both become equal to $[\phi_0 - m\tau d\phi(t)/dt]$. In practice, since high-order harmonics are the most disturbing, the audible distortion will arise mainly from modulating frequencies extending up to 1 kc/s, so that the equivalence may be taken as valid for $\tau \leqslant 100$ microsec or for path differences of up to about 30 km. This equivalence of the changes represented by eqns. (6) and (7) is referred to in Section 3.7.

DISCUSSION BEFORE THE ELECTRONICS AND COMMUNICATIONS SECTION, 4TH APRIL, 1960

Mr. G. Millington: The problem discussed in the paper forms good illustration of the difficulties so often imposed by the ropagation medium when a new system of communication is xploited. Frequency modulation has obvious attractions as ompared with amplitude modulation for v.h.f. sound broadasting, but I should like to ask whether in respect of multi-path istortion amplitude modulation is as susceptible as frequency odulation, and what bearing this may have on the choice of odulation for future applications.

Presumably tests were made to check that the delayed signal enerated by the simulator in Fig. 1 was itself undistorted, so at the distortion in the combined signal was genuinely due to e delay time. With regard to the subjective tests, I realize at it is difficult to give precise definitions of the different grades f distortion, but I feel that the criteria adopted are unsatisfactory. 'Perceptible' and 'disturbing' are general terms that need ualification, and I notice that in presenting the paper the author id, in fact, define more closely what he means by a perceptible s compared with a just perceptible annoyance.

With regard to the absolute tests, I am intrigued by the uggestion that the adjustment of a knob was adopted only or engineers since it required some skill to produce a desired hange. Surely it is a matter of experience that could be cquired by any intelligent person rather than of engineering kill as such.

It would have been useful if the author could have related his esults more closely to practical working by describing, for nstance, some of the situations in which multi-path distortion an be troublesome. I have the feeling that in most cases the ime delays are shorter than those mainly considered in the paper, nd it would have been interesting to have been told of the pecial topological conditions in which long time delays can ccur. In the case of reflections from aircraft, the time delays re usually short from the point of view of sound broadcasting, nd in the demonstration given the extremely unpleasant effect was almost certainly due to the Doppler change in frequency ather than to the time delay.

Mr. A. P. Hale: Would a narrower-band frequency-modulation ystem give better results regarding multi-path distortion without ny corresponding major disadvantages?

The use of directional aerials at the receiving site would give ery great improvement in results; examination of the graphs hows that even the use of an aerial of low directivity, such as n H-aerial, could result in the interference becoming imperceptible in very many cases.

Mr. H. Page: The original reason for using pre- and deemphasis in f.m. transmissions was to improve the signal/noise atio, and the time-constant was decided on this basis. Somewhat extravagant claims have been made (and are still sometimes repeated in published articles) for the improvement it is possible to obtain. In fact, the improvement realized in practice is very small for two reasons: first the average programme level must be reduced to avoid exceeding the peak deviation of the system; and, secondly, the ear acts as a natural de-emphasis network, and the subjective effect of the additional de-emphasis network used in the receiver is small.

However, as the author points out, de-emphasis has the beneficial effect of substantially reducing multi-path propagation distortion, and this may be the most important justification for its retention. If the author agrees with this conclusion, does he consider that the pre-emphasis time-constant of 50 microsec used in Europe is the best compromise?

Mr. F. C. McLean: In the imperfect receiver B, to what extent does the effect noted depend upon the signal input? Would the curve be different if shown for an input of 1 mV in comparison with 250 μV?

It is difficult to imagine that when there is a large difference of path between two signals, the indirect signal can be so very high compared with the direct signal. Are there any data giving the percentage of times at which various ratios between the wanted and unwanted fields are to be expected?

Mr. J. Moir: The distortion produced by long path reflections sounds remarkably similar to that produced by a rubbing voice coil and results in quite a few complaints, which should really be directed at the receiver manufacturers.

Have the B.B.C. listener research group any statistics showing the number of listeners that complain about the distortion introduced by long path reflections? Although the trouble can be very obvious with high-quality sound-reproducing equipment I suspect that many of the ordinary broadcast receivers have such a narrow bandwidth that they are more or less immune from trouble.

It has been suggested that a large amount of equipment is required to ensure that the aerial is pointing in the correct direction. However, if the incoming signal is greatly attenuated by some such simple device as inserting a piece of paper in the aerial input socket, the signal can generally be brought down below the level at which adequate limiter action is obtained. Under these conditions the direction and height at which to mount the aerial can be decided by a simple listening test. After clamping the aerial in this position the paper can be removed to increase the signal to the maximum obtainable with a fair degree of confidence that the aerial is pointing in the right direction for minimum distortion.

Are the expensive tuner units sold to the high-fidelity enthusiast any better in respect of this distortion than the commercial f.m. receivers sold at a much lower price?

Dr. G. J. Phillips: One question that has been raised concerns

the path differences that are significant in practice. Experience in the B.B.C. Research Department has shown that, when complaints of multi-path distortion arise, the path differences involved are usually in the range 8–24 km. For shorter path differences a large amplitude of reflected signal can be tolerated without distortion, so that trouble rarely occurs. Path differences exceeding 24 km are probably large compared with the average spacing of hills; screening of the receiver from the direct signal but not from the reflected signal is then less likely, and the extra path travelled greatly weakens the reflected signal.

One evening in 1957 I was able to hear multi-path distortion of a f.m. transmission from Sutton Coldfield as received in the London area. I believe it arose through reflection from an aurora; this is, of course, a rare occurrence and would in any case be detected only well outside the normal service area of a transmitter.

Mr. L. W. Turner: Regarding Mr. Moir's question c cerning the extent of multi-path trouble among v.h.f. listen it is difficult to arrive at an accurate figure for the percentage listeners affected, but in the first two or three years after start of the v.h.f. service in 1954 a significant flow of complai was received from listeners which, clearly, were concerned w multi-path distortion. It was, indeed, a matter of considera concern to the B.B.C., because the complaints were decide increasing in number as v.h.f. listening spread.

In the last two or three years, however, and coincident w the improvement in design of receivers—although, perhaps is disappointing that there has not been greater use of bet aerials—there has been a steady fall in the number of complai from listeners, despite the fact that there are now some 3 mill v.h.f. receivers in use whereas in the first two years there w perhaps less than a quarter of a million.

THE AUTHOR'S REPLY TO THE ABOVE DISCUSSION

Mr. R. V. Harvey (*in reply*): I agree with Mr. Millington as to the difficulty of specifying subjective criteria for distortion. The four criteria described in the paper may be regarded as labels for four adjacent sections of the region extending from 'imperceptible' to 'intolerable', and, judging by the results, their interpretation is very consistent in practice. Referring to the absolute tests, no suggestion was intended that only engineers were capable of adjusting knobs, but their patience and experience are undoubtedly required to assess the distortion on varying programmes and to make fine adjustments based on the average rather than the instantaneous annoyance.

In reply to Mr. McLean, the distortion heard from receiver B was indeed a function of the signal level in so far as the a.m. suppression ratio varies with signal level. The tests were conducted at a level at which the suppression ratio was the value given. Maintaining a good suppression ratio down to low signal levels is a matter of receiver design, and, in reply to Mr. Moir's suggestion, our experience has shown that an expensive tuner or receiver is not necessarily better in this respect than a cheap one.

I am grateful to Mr. Page for relating the chequered history of pre- and de-emphasis in f.m. broadcasting. The reduction in multi-path distortion that it affords is indeed very valuable with a time-constant of 50 microsec and this may well be the best compromise. The use of a longer time-constant would give further improvement in this respect but would begin to degrade the effective signal/noise ratio and increase the possibility of co-channel interference. Above a value of, say, 200 microsec, the accurate equalization of programme might become difficult and the receiver hum level could become appreciable. Part of the benefit of pre-emphasis arises as a result of the necessary reduction of the deviation, as Mr. Hale suggests, but the associated high-frequency accentuation avoids some of t disadvantages of a narrow-band f.m. system.

I thank Dr. Phillips for the information on the statistics of t incidence of delayed signals, and Mr. Turner for his comme on the incidence of the annoyance among listeners; th answered some of the questions raised by Messrs. McLe Millington and Moir. It is difficult to obtain exact statist of the annoyance caused by multi-path distortion, as m listeners are not able to distinguish this from other forms distortion or interference. Mr. Moir may be correct in su gesting that, in many instances, the complaints should directed at the receiver manufacturers, and here the B.B. Research Department has been doing what it can to dissemina any new ideas on simple and effective improvements in receiv design. I agree entirely with the comments of Messrs. Moir a Hale regarding the directivity and orientation of the aerial, b the implementation of these ideas is really the job of the servi man. However, manufacturers might well consider supplyi some simple form of room or loft aerial with instructions f its orientation.

Mr. Millington raises the question of the impact of multi-pa distortion on the choice between a.m. and f.m. for a broadca service. Though it is certainly true that an a.m. service wou be affected only by very large reflected signals of long dela I believe that the majority of complaints of distortion would n have arisen were it not for the relative freedom from noise a other forms of interference conferred by an f.m. system. T experience gained as a result of these complaints has, in fa led to improvements in receiver design which have reduced t number of dissatisfied listeners, as mentioned by Mr. Turne I am sure that, with further understanding and co-operatio this improvement will continue.

21.396.812.44

The Institution of Electrical Engineers
Paper No. 3285 E
Sept. 1960

SUNSPOT-CYCLE VARIATIONS IN THE DISCREPANCIES BETWEEN PREDICTED AND OBSERVED FREQUENCIES FOR USE IN RADIOCOMMUNICATION

By R. J. HITCHCOCK, M.A., G. O. EVANS, B.Sc., Associate Members, and R. NAISMITH, Member.

(*The paper was first received* 21*st January, and in revised form* 3*rd May,* 1960.)

SUMMARY

A marked variation occurs over the sunspot cycle in the discrepancies between predicted F2-4 000 km m.u.f. and the observed times of fade-in and fade-out on an 18·4 Mc/s Bombay–London circuit. This is particularly marked in summer and is attributed to the influence of the sporadic-E layer, although there is no marked sunspot-cycle effect attributed to this layer.

Any assessment of the accuracy of predictions analysed on the assumption of F-layer propagation must be considered in relation to the phase of the sunspot cycle to which it refers.

(1) INTRODUCTION

During 1953–55 considerable attention was focused on the discrepancies between the predicted standard* m.u.f. and the operational* m.u.f. as observed on radio circuits. In particular, attention was drawn to the magnitude of the summer error on medium- and long-distance east–west circuits on which, during the sunspot minimum years, it was possible between May and August to use frequencies considerably above those predicted by the 2-control-point method for the normal E and F2 layers only. It may not be generally realized that this major error associated with the summer condition is, in fact, at its greatest around the minimum of the sunspot cycle and that during the maximum of the cycle the predicted standard m.u.f.'s for these routes are, so far as the summer is concerned, generally in fair agreement with those derived from the observed fade-in and fade-out times on operational circuits. Since both Department of Scientific and Industrial Research (D.S.I.R.) predictions and circuit performance data are now available for a minimum and maximum of the sunspot cycle, it is possible for the first time to attempt to assess the accuracy of predictions as a function of sunspot activity. It must, however, be stressed that, owing to the nature of commercial operations, any such comparisons between the operational m.u.f., as defined by the monthly median of the daily fade-out times of a main day frequency and the predicted standard m.u.f., usually relate to only one point on the diurnal curve of standard m.u.f., and, as can be seen from Fig. 1, these fade-out times show a large change from sunspot maximum to minimum.

(2) OBSERVATIONS

Fig. 2 shows the differences between the operational and predicted standard m.u.f.'s as a percentage of the predicted standard m.u.f. at fade-out time for the 18·4 Mc/s frequency used on the Bombay–London circuit.

* These terms are defined in C.C.I.R. Recommendation 318, Document No. 573, Los Angeles, 1959.

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.
The paper is an official communication from the Radio Research Station of the Department of Scientific and Industrial Research.
Mr. Hitchcock is with Cable and Wireless, Ltd.
Mr. Evans is with the Post Office Engineering Department.
Mr. Naismith is at the D.S.I.R. Radio Research Station.

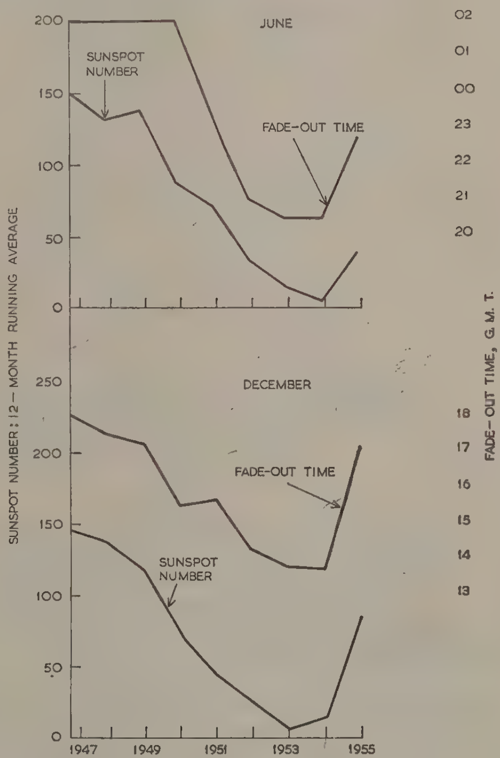


Fig. 1.—Variation of fade-out time of 18·4 Mc/s with sunspot number in June and December for the Bombay–London radiotelegraph circuit.

A negative percentage error indicates that the operational m.u.f. was greater than that predicted.

From the data given in Fig. 1 the curves shown in Fig. 3 have been prepared; the error has been averaged for the period May–August to represent the summer condition and for November–February to represent the winter condition. Fig. 3 also shows the average values of the 12-month running-average sunspot number for the summer and winter conditions.

It will be seen that the consistently large under-prediction associated with the summer months from 1950 to 1955 reached maximum proportions during and just after the lowest part of the minimum, i.e. 1954 and 1955, and that during the highest part of the maximum, namely 1947–48, the error appears to be very small. However, owing to the large difference between the fade-out times of 18·4 Mc/s at sunspot maximum and sunspot minimum, the accuracy of the prediction is—as has already been pointed out—being examined at times varying from approximately 2100 G.M.T. in June, 1954, to 0200 G.M.T. in June, 1947. The assumption that the 'summer error' virtually disappears

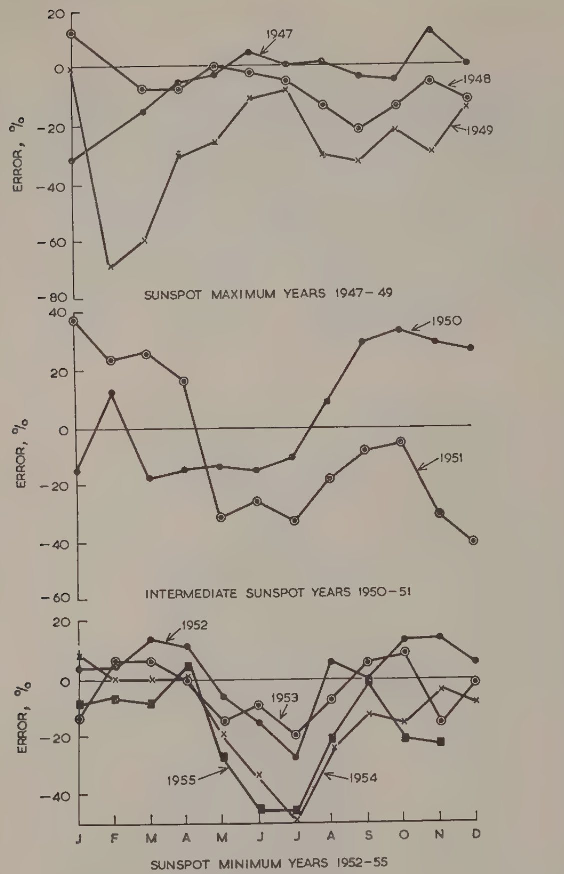


Fig. 2.—Average error in predicted monthly-median m.u.f. at fade-out time of 18·4 Mc/s at various parts of the sunspot cycle, for the Bombay–London radiotelegraph circuit.

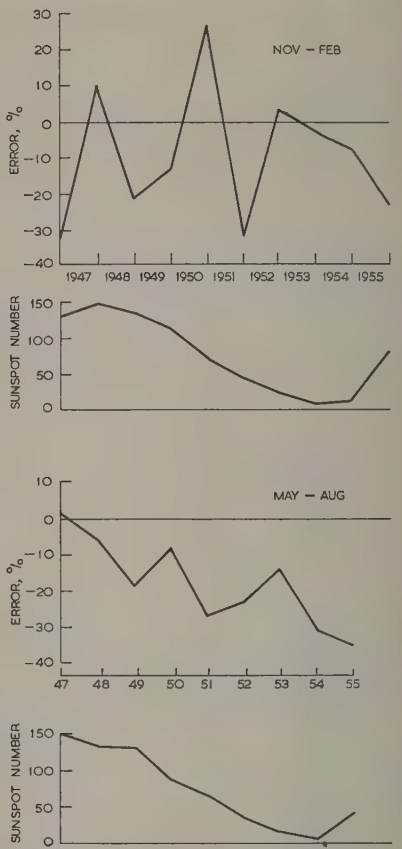


Fig. 3.—Average summer and winter errors in predicted monthly median m.u.f. at fade-out time of 18·4 Mc/s, for the Bombay London radiotelegraph circuit.

in sunspot-maximum years is therefore valid only if for a particular year the error is relatively constant between these hours. An examination of data recorded during the summer of 1956 (sunspot maximum) on the London–Colombo radiotelegraph circuit, which can be expected to have characteristics similar to those of the Bombay–London circuit, has shown, however, that the error in the m.u.f. prediction can, in fact, be assumed to remain fairly constant over the period 2100–0200 G.M.T.

Whereas this summer error, when it occurs, is consistently negative, errors during the winter, although often of considerable magnitude, appear to follow no particular trends.

No similar detailed analysis has been made for any other long-distance east–west route, but an examination of the general circuit 'post-mortem' information for circuits into London from Colombo, Shanghai, Taipeh, etc., showed, during summer months, the same decreasing correlation between predicted standard and operational m.u.f.'s as the sunspot minimum approached.

It might be concluded from this analysis that, during the summer at times of sunspot maximum, the assumption that propagation on these routes is governed by the normal E and F layers enables a reasonably accurate prediction of fade-out time to be made using the 2-control-point method. This conclusion does not, however, hold during summer at sunspot minimum, possibly owing to the effects of sporadic E-layer ionization.

(3) USE OF PREDICTIONS

The m.u.f. for the Bombay–London circuit is determined fo most of the time by conditions at the control point at the Londo end, and it should therefore follow fairly closely the changin values of fF2 measured at Slough when the 'summer error' i not present. (This assumption is basic to the following dis cussion.) If there were a period when the predicted values c fF2 for Slough were seriously in error, an opportunity would b afforded to assess the magnitude of this error from data on th performance of this particular circuit. Such a period occurre between July, 1950, and September, 1951, and is illustrated i Fig. 4, which shows noon values of fF2 measured at Sloug during the past 25 years. A fairly smooth envelope has bee drawn over the maxima and minima of the seasonal variation and this shows that the 1950–51 seasonal variation is con spicuously irregular. This irregularity was not anticipate and the predictions were consequently less accurate during th period. In this case the fF2-prediction error can be accuratel computed and is plotted in Fig. 5.

During the past 15 years more than half the predictions of fF for Slough noon have been within $\pm 7\%$ of the measured valu and it will be seen that this holds fairly well over the two samp years 1950–51 except for the winter period referred to abo when the error reached just over 40%. This error refers t noon values of fF2, but this determines the 4 000 km standar m.u.f. value given at times of fade-in and fade-out and ma

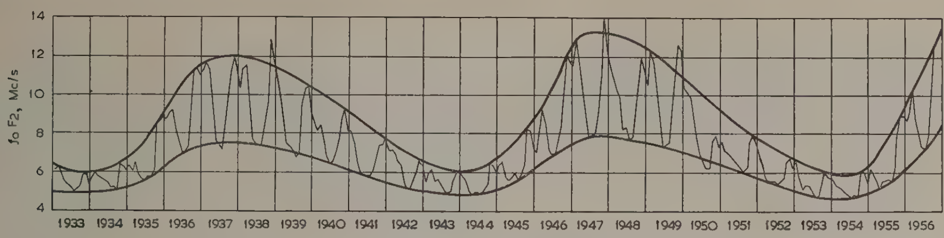


Fig. 4.—Measured noon values of foF2 at Slough during 1933–56.

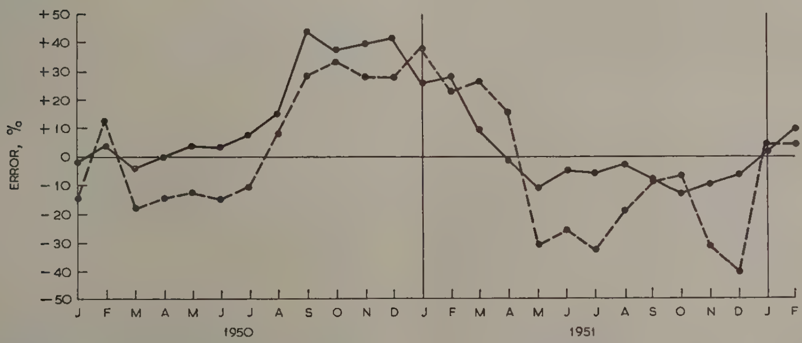


Fig. 5.—Average error in predicted noon values of fF2 at Slough and in predicted monthly-median m.u.f. for the Bombay–London path at fade-out time of 18·4 Mc/s during 1950–51.

—— Slough noon predictions.
---- Copied from Fig. 2.

therefore be compared with Fig. 2, which is redrawn on Fig. 5 for ease of comparison. If allowance is made for the 'summer error' (which is discussed below) the two graphs show quite clearly that the known error in prediction produced large discrepancies which were detected on this commercial circuit.

(4) THE SUMMER ERROR

The summer error is demonstrable only when propagation by the sporadic-E layer is not taken into account. The recognition in 1935* of the importance of the abnormal E-layer in facilitating long-distance high-frequency communication was followed by the introduction of this type of propagation into the forecasts issued by D.S.I.R. and by others of m.u.f.'s up to 2000 km range. In 1946 this fEs prediction was dropped, but those responsible for the maintenance of radiocommunication are fully aware of its importance. However, it has proved difficult to make accurate quantitative assessments of its influence on individual circuits.

In Section 2 reference is made to a hitherto unrecorded sunspot-cycle effect in the errors recorded on the Bombay–London circuit in summer when F2-layer propagation only is considered. This sunspot-cycle effect provides an opportunity of demonstrating the improvement which could result from the inclusion of a suitable forecast of propagation by the Es layer. Since there is no marked sunspot-cycle effect in the measured values of fEs, some further discussion of this quantity is required.

(5) INTERPRETATION

The Slough measurements of fEs and fF2 have been converted to values of standard m.u.f. corresponding to the distances of 2000 and 4000 km respectively, which are those normally associated with one reflection in the layer to which they refer. On the Bombay–London circuit the path is nearly due east, and the two control points, one in the E layer and one in the F2 layer, are approximately 1000 km (or 15° in longitude) apart. Since this corresponds to one hour in local mean time (L.M.T.) it is better to plot the diurnal variations at the two control points on the same diagram in terms of Greenwich mean time, and this has been done in Fig. 6. Furthermore, each quantity has been averaged over the period May–June to give a representative summer value for comparison with the circuit data described above.

The graphs show quite clearly that a frequency of 18·4 Mc/s would be controlled by the F2 layer at fade-out time and by either layer at fade-in time at sunspot maximum (1947) [Fig. 6(*a*)]. At sunspot minimum (1954) [Fig. 6(*c*)] the Es-layer controls the circuit both at fade-in and at fade-out times. Intermediately (1951), Fig. 6(*b*) shows that either layer may control at fade-out but the Es layer would still control at fade-in time. Thus, it is to be expected that at sunspot minimum a considerable error will appear in the predicted fade-in and fade-out times when the predictions are based simply on the normal E and F layers. At sunspot maximum, however, when the F2 layer controls, the error could be expected merely to reflect the error in the prediction of foF2, which, as has been mentioned in Section 3, is of the order of $\pm 7\%$.

The observations referred to in Section 2 show that this is, in fact, what happens, the error during the summer months being a maximum at sunspot minimum. Any assessment of the accuracy of m.u.f. predictions which takes account of the normal E and F layers only must therefore be considered in relation to the particular phase of the sunspot cycle to which the assessment refers.

* Appleton, E. V., and Naismith, R.: 'Some Further Measurements of Upper Atmosphere Ionization', *Proceedings of the Royal Society*, A, 1935, 150, p. 691.

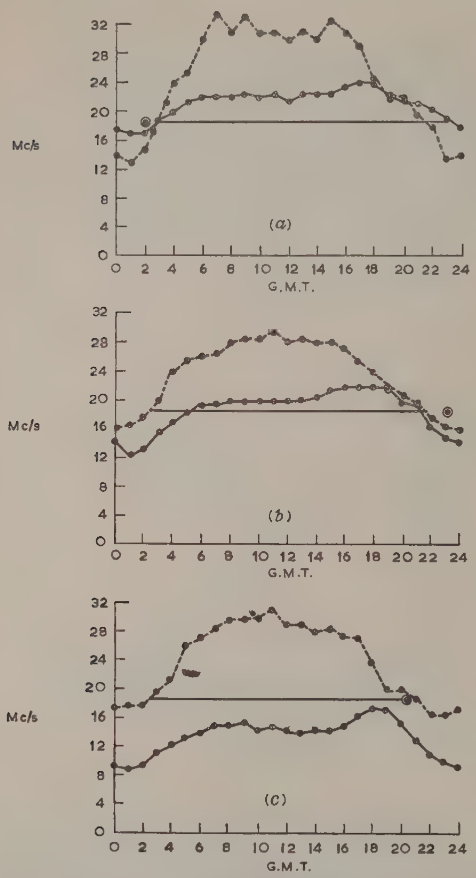


Fig. 6.—F2-4 000 km m.u.f. and Es-2 000 km m.u.f. derived from foF2 and fEs measured at Slough.

—— F2.
---- Es.
⊙ Actual fade-out times for June.
(*a*) 1947.
(*b*) 1951.
(*c*) 1954.

The actual fade-out times for June shown in Fig. 1 have been plotted on the average summer curves of Fig. 6, and it will be seen that the predicted fade-out time is in fairly close agreement with the actual fade-out time if both the Es and F2 layers a taken into account.

(6) FURTHER DISCUSSION

On a number of occasions the standard m.u.f.'s for the I and F2 layers are similar, and this will generally produce son extension of time during which the circuit can be operated. Fo example [see Fig. 6(*b*)], although 18·4 Mc/s may only be prop gated by the Es layer for, say, 40% of the time between 220 and 2315 G.M.T., it may be possible for the F2 layer to continu the propagation for a further 10% since the relationship of hig values of fEs and fF2 appears to be random.

On the other hand, at sunspot minimum, when only one laye is effective, the fade-out time should be nearer to that computed- as, in fact, it is in Fig. 6(*c*).

In Fig. 5 the discrepancy between the predictions for Novemb and December, 1951, and the observations is greater than th average referred to in Section 3. We have studied this poi and find that this error cannot be explained in terms of E values observed at Slough.

The prediction error for fF2 was near the average and th error due to the form of presentation was not greater than 10% Although both errors were in the direction shown, they do no account for half the error observed but further examinatio of the cause is outside the scope of the paper.

(7) CONCLUSION

The analysis of the discrepancies between F2-layer prediction and the performance of a Bombay–London circuit show tha inaccuracies in the basic prediction can be detected. It als shows the existence of a 'summer error' in the standard m.u.f prediction which varies over the sunspot cycle and makes i necessary to refer to the phase of the sunspot cycle in any discus sion of these discrepancies.

The 'summer error' is shown to be most likely due to propaga tion by the sporadic-E layer.

(8) ACKNOWLEDGMENTS

The paper is published by permission of the Engineer-in-Chie of Cable and Wireless, Ltd., the Engineer-in-Chief of the Pos Office, and the Director of Radio Research of the Departmen of Scientific and Industrial Research. The work described i the paper formed part of the programme of the Radio Researc Board.

621.317.784

The Institution of Electrical Engineers
Paper No. 3298 E
Sept. 1960
©

RESISTIVE-FILM MILLIWATTMETERS FOR THE FREQUENCY BANDS 8·2–12·4 Gc/s, 12·4–18 Gc/s and 26·5–40 Gc/s

By I. LEMCO, B.Sc., and B. ROGAL, B.Sc.(Eng.), Associate Members.

(*The paper was first received* 25*th March, and in revised form* 19*th May*, 1960.)

SUMMARY

A commercially developed version of a wide-band microwave milliwattmeter for the frequency band 8·2–12·4 Gc/s, based on Lane's principle of mounting a narrow resistive strip in the transverse plane of a waveguide, is described. The instrument gives power measurements in the range 1–100 mW with an accuracy within ±2%. Effects of temperature are discussed, and a robust version of the milliwattmeter is mentioned. Two models designed on the same principle for the frequency bands 12·4–18 Gc/s and 26·5–40 Gc/s are also described.

Comparison of these with other power-measuring instruments indicates that the resistive strip mounts are of high accuracy at these frequencies. This is obtained by the simplicity of design of these mounts resulting in lower power losses.

(1) INTRODUCTION

For some time there has been need for a relatively simple and accurate method of measuring microwave power levels in the milliwatt region. It has been shown[1] that errors of up to 10% can occur using thermistor or bolometer mounts at wavelengths of 3 cm, and these errors are considerably larger at higher frequencies.[2] Lane[3] has shown that a narrow strip resistive-film bolometer mounted symmetrically in the transverse plane of a rectangular waveguide with a short-circuiting termination at a distance of about $\lambda_g/4$ behind it will act as a non-reflecting power-absorbing device. The films used by Lane consisted of platinum deposits sputtered on to mica strips. It was found that if the d.c. resistance of the strip was made about equal to the wave impedance of the guide, $377\lambda_g/\lambda$ ohms, the slight inductive component due to the geometry of the strip could be cancelled out by adjusting the position of the reflecting termination to a distance slightly greater than $\lambda_g/4$. For a strip mounted in this fashion the high-frequency current distribution is practically the same as that for the d.c. case. Hence the device can be calibrated in terms of d.c. power dissipation. By attaching a thermocouple junction to the centre of the film an indication of temperature rise may be obtained by use of a galvanometer or other sensitive d.c. measuring instrument.

(2) DESIGN OF MILLIWATTMETER FOR THE FREQUENCY BAND 8·2–12·4 Gc/s

The construction of a milliwattmeter based on these principles and designed for commercial use in the frequency band 8·2–12·4 Gc/s is illustrated in Fig. 1. The film is deposited on to a 0·004 in thick crown-glass strip using a vacuum evaporation process. The plain glass strip is first mounted in a vacuum evaporation plant and a central length of 0·4 in is masked. A film of gold is then deposited on to the glass. The mask is removed and an evaporation of Nichrome is performed so that the d.c. resistance between the gold terminals is between 480 and 520 ohms. The resistances are based on the results obtained by Lane, and also provide a reasonable degree of tolerance for the evaporation process. Finally a protective coating of silicon monoxide is deposited on to the central portion of the strip, leaving a small length free at each end for electrical contact to the gold deposit. Films produced by the technique of sputtering platinum on to mica were investigated, but the resistances obtained by this process varied considerably, differences of

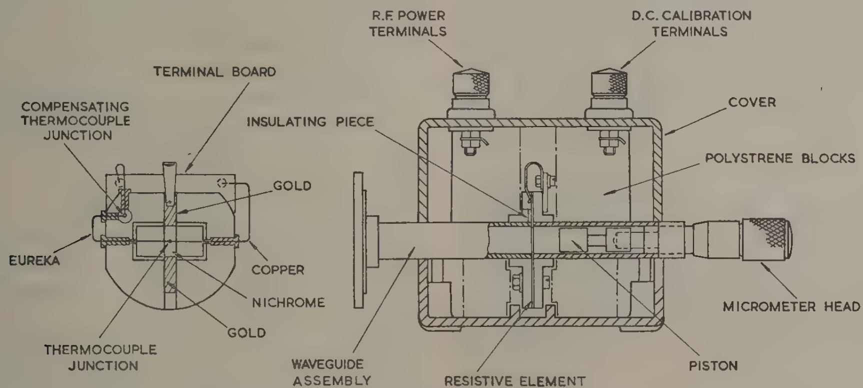


Fig. 1.—Construction of 8·2–12·4 Gc/s milliwattmeter.

up to 5 : 1 being noted among strips prepared under identical conditions. Glass was substituted for mica, as this is not a suitable substrate for the evaporation method, and Nichrome was used as the resistive medium in preference to platinum for economic reasons.

The finished strip is mounted in waveguide No. 16 (0·9 × 0·4 in i.d.) between two machined flanges, so that the only major gaps in the guide are 0·005 in wide slots where the glass emerges

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.

Mr. Lemco and Mr. Rogal are with Wayne Kerr Laboratories Ltd.

from the walls. The resistive element is connected directly to the guide at one end by one of the gold terminations. One of the pair of terminals marked 'D.C. Calibrate' is also connected directly to the guide. The other gold termination is held insulated from the guide, as far as direct current is concerned, and a connection is taken from it to the remaining terminal of the 'D.C. Calibrate' pair. This pair of terminals is used in the calibration of the instrument.

A copper-Eureka thermocouple junction is attached to the centre of the resistive film by a spot of adhesive. The thermocouple wires are taken through the centres of the side walls of the guide, the Eureka wire then being connected to a second copper wire so as to form another thermocouple junction. This junction is embedded within the flange assembly and provides a measure of temperature compensation for changes in ambient temperature. The two copper leads from the junctions are taken up to the pair of terminals on the outer cover marked 'R.F. Power'.

The short-circuiting plunger at the rear of the film is of a non-contacting cylindrical design. It consists of a series of low- and high-impedance sections each equal to $\lambda/4$ at midband. This type of plunger acts as a good short-circuit over a wide band. The lengths of the high- and low-impedance sections may vary from $0\cdot1\lambda$ to $0\cdot4\lambda$ before a deterioration in performance is observed. Also, the design has the advantage of being free from the effects of the relatively poor short-circuit at the rear of the plunger, where it is mounted in the end of the guide.

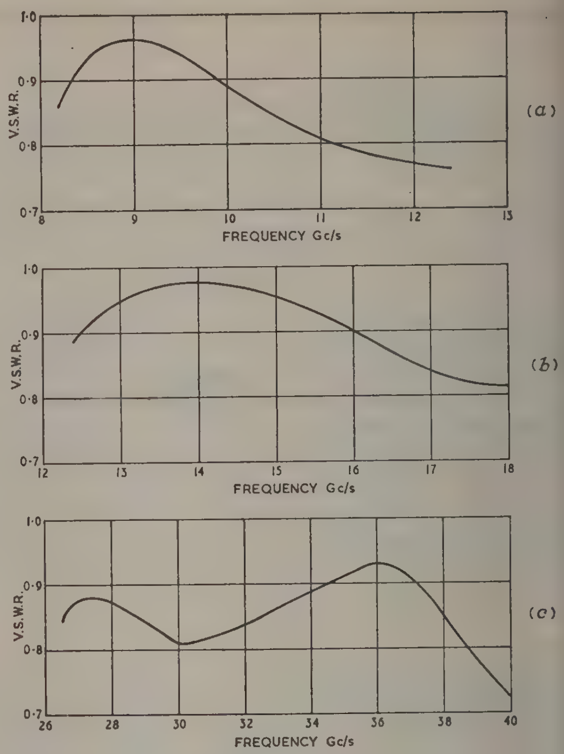


Fig. 2.—Variation of input v.s.w.r. with frequency.
(*a*) 8·2–12·4 Gc/s milliwattmeter.
(*b*) 12·4–18 Gc/s milliwattmeter.
(*c*) 26·5–40 Gc/s milliwattmeter.

(3) CALIBRATION AND PERFORMANCE

(3.1) Calibration and Accuracy

Using high-grade d.c. instruments the calibration may be performed with errors not exceeding ½%. During the calibration the 'R.F. Power' terminals are connected to a galvanometer. For powers up to 100 mW a linear relationship holds between power and galvanometer deflection.

A minimum sensitivity of 10 μV/mW was aimed at, but some models have been found to have sensitivities as high as 15 μV/mW. The sensitivity performance depends on the nature of the intimate contact between the thermocouple and the film, and once a firm contact has been made the performance is not critical with respect to time. Using a galvanometer of sensitivity $0\cdot5$ mm/μV with an internal resistance of 100 ohms, the time-constant of the milliwattmeter was 15 sec. For measurements of powers below a milliwatt a more sensitive galvanometer or d.c. indicating device may be used.

Fig. 2(*a*) shows the variation of v.s.w.r. with frequency for optimum micrometer setting. Measurements on a large number of models have shown that the v.s.w.r. tends to be better at the high-frequency end of the range for lower values of the film resistance, i.e. values nearer to 480 ohms than to 520 ohms. This is to be expected from consideration of the formula for wave impedance. In use a correction for power loss due to reflection from the film can easily be applied. For a v.s.w.r. of $0\cdot8$ this is just over 1%.

Three of the production models of the instrument have been checked for accuracy against power-measuring standards such as the double-vane force-operated wattmeter[4] and water calorimeters. These comparisons have been carried out at two different establishments with independent equipments. In both establishments power levels were set up in a main guide using the force-operated wattmeter and a water calorimeter. The milliwattmeter was attached to the main guide by a calibrated directional coupler. The deviation in measurement of power between the milliwattmeter and the other instruments has always been less than 2%. This confirms Lane's results.[3]

(3.2) Temperature Effects

Experiments were carried out on the effect of temperature changes on the resistance of the film. With the instruments placed in a thermostatically controlled oven and with a steady direct current passed through the film, the voltage drop across the film was noted as the temperature was raised from 15° to 70° C. Any change in voltage was found to be less than 1%.

The deflection of a galvanometer connected to the 'R.F. Power' terminals was also noted. With a current passing through the film corresponding to a power dissipation of 6 mW, changes of 5% of the galvanometer deflection were observed. With the current switched off the galvanometer still indicated a small deflection at some temperatures. This is due to differential thermo-electric voltages between the two thermocouples and also to additional e.m.f.'s set up at the connections between the galvanometer leads and the 'R.F. Power' terminals. Since the resistance of the film does not alter, these small e.m.f.'s must be largely responsible for the changes in deflection with the current on. The effects are difficult to eliminate, but there has been a tendency over very long periods of stable temperature conditions for these small voltages to decrease in magnitude. In practice, for powers up to 100 mW the percentage error will decrease as it is not a function of the power dissipation in the film. Even in the 1–10 mW region, errors of less than 3% would be obtained up to a temperature of 40° C by using the instrument calibrated at 20° C. Owing to the nature of the construction of the oven it was impossible to check the microwave performance with temperature. Since the d.c. resistance does not vary one would expect the input v.s.w.r. to remain constant. This was indeed observed by measuring the v.s.w.r. of the instrument

quickly after it had been withdrawn from the oven at a temperature of 70° C.

(4) A ROBUST MILLIWATTMETER

For strenuous or field applications the use of a glass vane may not be acceptable or convenient. A more robust form of the milliwattmeter has been developed for a special purpose and for single-frequency operation. In this instrument the glass vane is replaced by a piece of resistance card. Standard resistance card of 200 ohms per square has been found to be most suitable, since a piece of about the same height and width as the glass vane has a resistance near 500 ohms. The card is mounted as in the glass version, electrical contact being made to each end by depositing silver from a solution of silver in alcohol. The required length of uncoated card is obtained by using a mask in a somewhat analogous manner to the glass-vane case. The tunable piston is replaced by a soldered short-circuit at the correct distance behind the strip.

Measurements on this type of milliwattmeter with a single thermocouple junction attached to the film indicate that the sensitivity is only about 70% of that of the glass-vane type. This is probably due to the relatively thick (0·014 in) card, which absorbs some of the heat. To overcome this, two thermocouples in series have been attached to the centre of the strip, one above the other but not in contact. Using this arrangement the sensitivity has been effectively doubled. With the same galvanometer as used earlier the time-constant of this mount was measured as 22 sec. This is higher than with the glass vane owing to the longer period required to heat the thermal mass of the card. Comparison of this type of milliwattmeter with the glass-vane instrument has shown an equivalent accuracy.

Experiments have been carried out using carbon films deposited on a Bakelite base 0·005 in thick. This device in conjunction with a single thermocouple had a time-constant of 17 sec and a sensitivity similar to that of the glass-vane instrument.

Temperature measurements similar to those described in Section 3.2 were carried out on the card milliwattmeter. The resistance was found to vary with temperature, a maximum change of 4% being noted at 70° C. In terms of v.s.w.r. this effect is not appreciable, and hence the accuracy is not greatly affected.

(5) MILLIWATTMETERS FOR HIGHER FREQUENCIES

(5.1) 12·4–18 Gc/s Milliwattmeter

A milliwattmeter for the frequency band 12·4–18 Gc/s has been designed. This mount is constructed in waveguide No.18 (0·622 × 0·311 in i.d.) and is substantially a scaled-down version of the lower-frequency model, as shown in Fig. 3. The vane is narrower and the resistive film across the guide is shorter to conform with the lower height of the guide. The construction is very similar, and since the various holes in the guide for thermocouple wires and the vane are of much the same size as in the lower-frequency case, larger power losses may be expected here. The non-contacting short-circuiting plunger may also be expected to have larger losses at these frequencies. The losses were investigated by measuring the small v.s.w.r. of the mount with the resistive element replaced by a plain glass vane. The method is that of Roberts and Von Hippel.[5] Table 1 shows the results obtained at three frequencies in the band. The highest loss is 5% and the loss is smaller at mid-band.

Fig. 2(*b*) shows the v.s.w.r. performance of the completed mount for optimum micrometer settings. It will be seen that the v.s.w.r. does not fall below 0·8. The d.c. resistance of the vane was 495 ohms. The calculated values of the wave impedance in this waveguide for the full frequency range vary from 440 to 610 ohms.

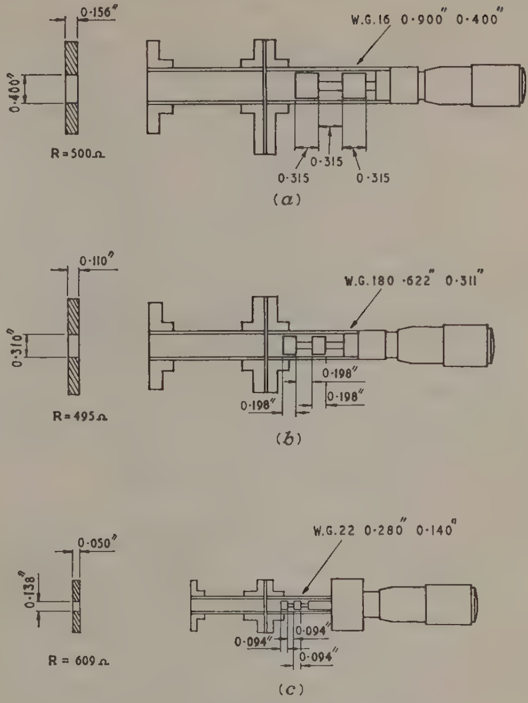


Fig. 3.—Comparison of dimensions.

(*a*) 8·2–12·4 Gc/s milliwattmeter.
(*b*) 12·4–18 Gc/s milliwattmeter.
(*c*) 26·5–40 Gc/s milliwattmeter.

Table 1

LOSSES IN 12·4–18 Gc/s MILLIWATTMETER

Frequency	Power loss, percentage of incident power not reflected
Gc/s	%
13·0	5·0
15·0	3·2
17·0	4·9

The sensitivity performance is slightly better than that of the lower-frequency instrument. Since the vane in this case is of smaller dimensions the final stable temperature reached for the same input power will be higher than in the case of the larger instrument.

A comparison of the performance of this milliwattmeter was made against an ordinary commercial thermistor mount. Power levels were first set up on a bench using the resistive-film instrument. Then the milliwattmeter was removed and the bench terminated by the thermistor mount, which was connected to a measuring bridge. Table 2 shows the power level indicated by each instrument at the same frequency. Corrections for the v.s.w.r. of the thermistor mount were made. Since the losses in the milliwattmeter mount are modified by the presence of the resistive film, no correction for these could be made with certainty. Even so, it will be seen that the milliwattmeter shows a higher power level at most frequencies, indicating a higher

Table 2

COMPARISON OF 12·4–18 Gc/s MILLIWATTMETER WITH A THERMISTOR MOUNT

Frequency	Power measured	
	12·4–18 Gc/s milliwattmeter	Thermistor mount
Gc/s	mW	mW
13·0	3·00	3·00
14·0	3·00	2·96
15·0	3·00	2·25
16·0	3·00	2·77
17·0	4·00	3·88
18·0	4·00	3·75

accuracy than that of the thermistor mount. It is intended to compare the performance of the milliwattmeter with that of a more reliable device such as an enthrakometer[6] or a water calorimeter.

(5.2) 26·5–40 Gc/s Milliwattmeter

A prototype milliwattmeter for the band 26·5–40 Gc/s has also been designed. The mount is constructed in waveguide No. 22 ($0·280 \times 0·140$ in i.d.) and is again a further scaling-down of the sizes used in the lower-frequency cases. The dimensions are given in Fig. 3. The glass vane in this case would have been so narrow, 0·050 in, that it was impossible to manufacture. Instead, an attempt was made to deposit the required narrow resistive strip on to a glass piece whose width was 0·11 in, the same as for the 2 cm milliwattmeter. This has failed owing to the difficulties involved in the technique of evaporating the film on to the glass and keeping it within desired resistance limits. In future it is hoped to deposit a carbon layer on to a 0·004 in Bakelite strip, as described in Section 4, and experiments into the feasibility of this construction are under way. Meanwhile, an ordinary strip of resistive card has been cut down to the required width from 200-ohms-per-square sheet. This has been treated and mounted in the waveguide as described earlier. To keep losses low, the card has been pared off at the rear non-resistive side so that the overall thickness is 0·005 in. Fig. 2(*c*) illustrates the variation of v.s.w.r. with frequency for optimum micrometer setting. This falls below 0·8 only at the extremes of the band. For this frequency range the wave impedance of the guide varies from 450 to 660 ohms, and the d.c. resistance of the element in the guide is 609 ohms. When the mount using a resistive film mounted on a Bakelite strip is constructed it will be compared with the other power-measuring instruments. A comparison of the existing mount with a thermistor mount has been made, and this shows an equivalent accuracy over most of the band. There are discrepancies at the ends of the band, but it is hoped to eliminate these with the new strip. The results are given in Table 3. The thermistor mount was fully corrected for v.s.w.r. and power losses.

Table 3

COMPARISON OF 26·5–40 Gc/s MILLIWATTMETER WITH A THERMISTOR MOUNT

Frequency	Power measured	
	26·5–40 Gc/s milliwattmeter	Thermistor mount
Gc/s	mW	mW
26·6	3·06	3·21
28·6	3·04	2·83
31·4	2·06	1·99
33·0	3·05	2·87
35·8	3·04	3·26
38·9	3·07	3·88

(6) CONCLUSIONS

The developments described have confirmed Lane's results for 3 cm wavelengths and have shown that the frequency range can easily be extended to cover 8·2–12·4 Gc/s. Results obtained with the resistive-film milliwattmeters indicate that this method of measurement of microwave power is superior to the normally used thermistor-mount methods both in accuracy and in ease of use and calibration for wide-band operation. Although an instrument for the frequency band 18–26·5 Gc/s has not been described here, a suitable design based on the principles given above could be made if required.

(7) ACKNOWLEDGMENTS

Thanks are expressed to Mr. J. A. Lane of the Radio Research Station, D.S.I.R., for some useful discussion on the subject, and to the Directors of The Wayne Kerr Laboratories Ltd., for permission to publish the paper.

(8) REFERENCES

(1) LANE, J. A.: 'The Measurement of Power at a Wavelength of 3 cm by Thermistors and Bolometers', *Proceedings I.E.E.*, Paper No. 1918 R, November, 1955 (**102 B**, p. 819).
(2) MONTGOMERY, C. G.: 'Technique of Microwave Measurements' (M.I.T. Radiation Laboratory Series, McGraw-Hill, New York, 1947), p. 154.
(3) LANE, J. A.: 'Transverse Film Bolometers for the Measurement of Power in Rectangular Waveguides', *Proceedings I.E.E.*, Paper No. 2488 R, January, 1958 (**105 B**, p. 77).
(4) CULLEN, A. L., ROGAL, B., and OKAMURA, S.: 'A Wide-Band Double-Vane Torque-Operated Wattmeter for 3 cm Microwaves', *Transactions of the Institute of Radio Engineers*, April, 1958, **MTT–6**, p. 133.
(5) BARLOW and CULLEN: 'Microwave Measurements' (Constable, 1950), p. 123.
(6) COLLARD, J.: 'The Enthrakometer, an Instrument for the Measurement of Power in Rectangular Waveguides', *Journal I.E.E.*, 1946, 93, Part IIIA, p. 1399.

21.396.677.832

The Institution of Electrical Engineers
Paper No. 3303 E
Sept. 1960

A SLOT-EXCITED CORNER REFLECTOR FOR USE IN BAND V (610–960 Mc/s)

By D. J. WHYTHE, B.Sc.(Eng.), Associate Member, and K. W. T. HUGHES, B.Sc.

(*The paper was first received* 24*th February, and in revised form* 31*st May*, 1960.)

SUMMARY

The directional aerial described employs a corner reflector excited by a four-tier slotted cylinder. It was used during a series of test transmissions in Band V (610–960 Mc/s) which required an aerial capable of withstanding severe climatic conditions for long periods without attention.

A design suitable for operation at 900 Mc/s is described; the measured results include admittance characteristics and radiation patterns in the horizontal and principal vertical planes.

(1) INTRODUCTION

The paper describes a transmitting aerial which was developed for propagation tests employing horizontal polarization in Band V (610–960 Mc/s). The tests were conducted by radiating from a fixed site towards a number of receiving stations arranged along approximately the same radial from the transmitting station. To simplify the selection of receiving sites, the beam-width of the transmitting aerial in the horizontal plane was required to extend approximately 40° between the half-power points. The aerial was required to withstand rigorous climatic conditions (including snow and ice formation) for long periods without inspection.

Other features of the aerial specification are summarized below.

(*a*) The power gain to be not less than 15 dB relative to a half-wave dipole. (This implies a half-power beamwidth in the vertical plane of about 12° in order to satisfy the beamwidth requirement in the horizontal plane.)

(*b*) The standing-wave ratio of the aerial referred to 70-ohm coaxial feeder to be better than 0·9 at the specified frequency.

(*c*) The aerial to be capable of radiating the output of a transmitter which can be modulated either with square waves having a mean power of 50 watts, or with 0·5 microsec pulses having a repetition frequency of 1 kc/s and a peak power of 10 kW.

(*d*) The performance to be unaffected by the presence of the supporting mast and to show a minimum of change due to snow and ice formation.

The aerial to be described fulfilled this specification at 900 Mc/s.

(2) THE DESIGN OF THE AERIAL

(2.1) Preliminary Considerations

The required directivity could have been achieved with an array of Yagi aerials, but this arrangement was not used because of difficulties of feeding and de-icing. It was decided instead to use an aerial incorporating a reflecting surface which, with a polythene cover over the aperture, gives protection from the weather. A plane-sided reflector was adopted; parabolic and cylindrical reflectors were considered, but they would have been more difficult to manufacture and they appeared to offer no great advantage.

When a plane-sided, or corner, reflector is excited by conventional electric dipoles, it is usual to orient the reflector apex parallel to the axis of the dipoles.[1] Since horizontally polarized waves were specified, this form of excitation would have required mounting the reflector apex horizontally. With this arrangement, however, the required vertical directivity could not be conveniently achieved. It was therefore decided to mount the reflector with its apex vertical and to excite it with the equivalent of an array of magnetic dipoles—a slotted cylinder. With this arrangement it was found experimentally that the requirements could be satisfied by a corner reflector $4 \cdot 6\lambda$ high and $1 \cdot 17\lambda$ wide.

(2.2) Description of the Aerial

An outline drawing of the aerial with its polythene cover removed is shown in Fig. 1. It consists of a 45° corner reflector

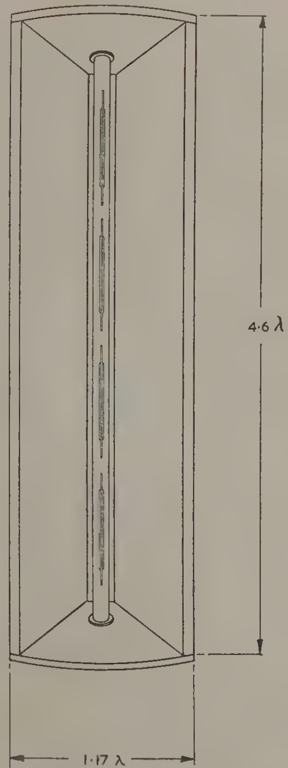


Fig. 1.—Aerial with cover removed.

with closed ends which support the exciting element. To ease the problem of mounting the aerial on a mast the apex of the corner is truncated; this has been found[1] to have a negligible effect on the radiation pattern.

For maximum power gain, the slotted cylinder should extend over the full length of the corner reflector, and the voltage across the slot should be constant in amplitude and of the same

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.

Mr. Whythe is with the British Broadcasting Corporation.

Mr. Hughes, formerly with the B.B.C., is now with Racal Engineering Ltd.

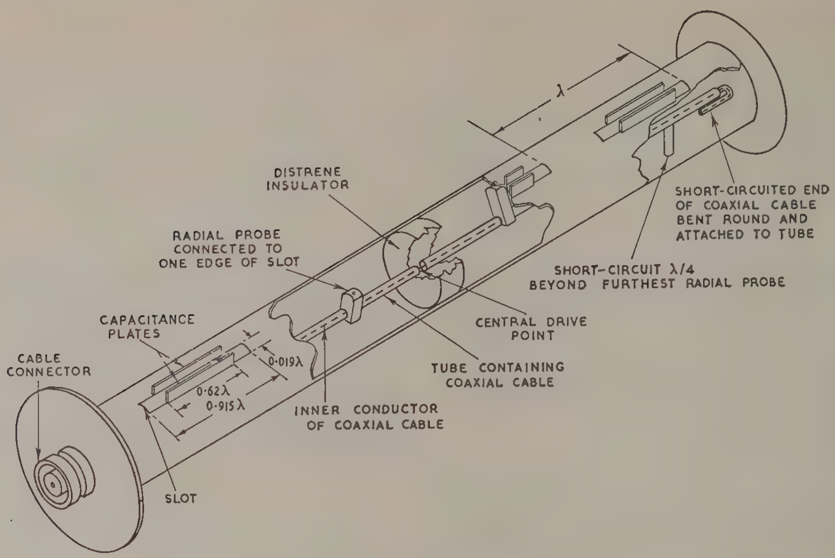


Fig. 2.—The slotted cylinder.

phase along its length. A good approximation to this condition may be obtained by dividing the slot into a number of sections driven equally at their mid-points.

A slotted cylinder can be considered as a balanced transmission line loaded by distributed shunt inductance.[2] It behaves in a manner similar to a waveguide, with a critical (or cut-off) frequency. The effect of the distributed shunt inductance may be modified and controlled by the addition of capacitance plates placed along the length of the slot. In this way a centre-driven slot of any length may be made to have an electrical length corresponding to half-wave resonance. Each slot used in the aerial is $0\cdot915\lambda$ long; four such slots are used to excite the reflector. Capacitance plates are fitted which extend over a length of $0\cdot31\lambda$ on each side of the central drive-points.

Fig. 2 shows the general arrangement of the slotted cylinder. It contains a tube which is split at its mid-point and which itself contains a coaxial cable, shorn of its outer conductor. The generator is effectively connected across the gap at the mid-point of the inner conductor, through a series compensating stub to be described later. Equal co-phased slot voltages are established by radial tapping probes attached to the tube at intervals of one wavelength, the probes being connected to opposite sides of the slots in the upper and lower halves of the aerial.

The tube is supported by Distrene spacers; its ends are terminated by short-circuits to the inner wall of the cylinder $\lambda/4$ beyond the outermost probes. The cable inside the tube is connected to the feeder by a socket at the lower end of the cylinder. The inner conductor of the cable is not connected directly to the further half of the tube but is continued across the break to the end of the cylinder to form a series-impedance compensating stub. The admittance compensation of this arrangement is analogous to the impedance compensation of a shunt-connected stub applied to a half-wave dipole; its use also has the advantage that no soldered electrical connection is required at the mid-point of the inner tube. The end of the compensating stub protrudes from the end of the cylinder, where it is readily accessible, and from this point it continues with its original outer braiding. It is terminated in a short-circuit; its length is chosen to give the required overall reactance compensation and a good overall admittance match. Intermediate admittance correction is achieved by adding shunt capacitance to the inner conductor on each side of the centre spacer, in the form of cylindrical sleeves. The equivalent circuit is shown in Fig. 3.

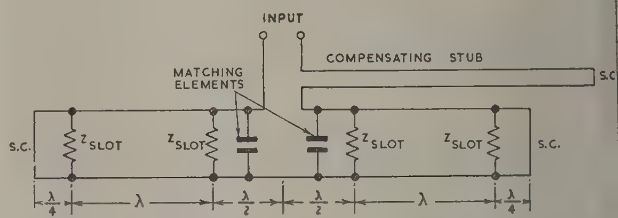


Fig. 3.—Equivalent circuit of the feeder system.

The coaxial cable used is Uniradio No. 32; this limits the mean power-handling capacity of the aerial to 70 watts at 900 Mc/s. A simple modification would enable Uniradio No. 1 to be used whereby the mean power rating could be increased to 230 watts.

The slotted cylinder and the metal components of the feeder system are silver plated. The corner reflector is made of a weather-resisting aluminium alloy. A sheet of black polythene, $\frac{1}{8}$ in thick, is secured over the aperture to weatherproof the interior.

(3) RADIATION PATTERNS AND GAIN

(3.1) Horizontal Radiation Pattern

Preliminary horizontal radiation pattern (h.r.p.) measurements were made using a single slot in an open-ended corner reflector. Experiments were carried out to determine:

(a) The required angle between the sides of the reflector.
(b) The side dimensions.
(c) A suitable range of spacings between the slotted cylinder and the apex of the reflector (the reflector was not truncated at this stage).

These parameters were found to be interdependent to some extent, but a set of values was arrived at for which the dimensions were not critical.

Measurements of h.r.p. were made with the slot both facing the apex and facing the aperture. As a result of these preliminary tests it was decided to use a slotted cylinder of $0 \cdot 14\lambda$ diameter, mounted with its slot facing the aperture of a reflector having sides $1 \cdot 25\lambda$ wide and an included angle of 45°. The h.r.p. was found to show little change if the spacing between the axis of the cylinder and the apex of the reflector was varied between $0 \cdot 4\lambda$ and $0 \cdot 6\lambda$. The final position of the cylinder within the reflector was determined later as a result of admittance measurement. The h.r.p. ultimately achieved is shown in Fig. 4.

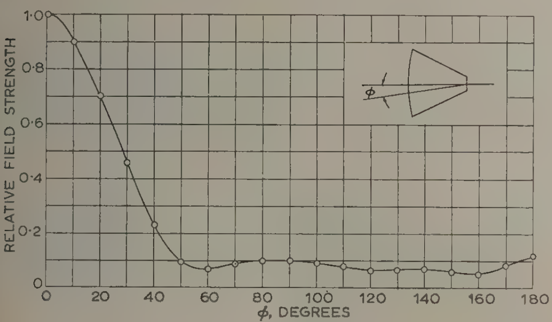


Fig. 4.—Horizontal radiation pattern of the 900 Mc/s aerial.

(3.2) Vertical Radiation Pattern

The vertical radiation pattern (v.r.p.) in the forward direction is very similar to that of an array of collinear half-wave dipoles having the same spacing as the slots. It was found that the original specification could be satisfied by four slots with their centres spaced $1 \cdot 0\lambda$ apart. Fig. 5 shows the v.r.p. measured

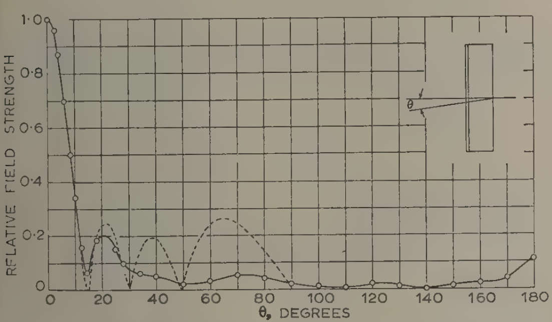


Fig. 5.—Vertical radiation pattern of the 900 Mc/s aerial.
o—o—o Measured v.r.p.
------ Theoretical v.r.p. of four collinear $\lambda/2$ dipoles spaced at $1 \cdot 0\lambda$ ($0 < \theta < 90°$).

at 900 Mc/s; for comparison the theoretical pattern for four collinear half-wave dipoles with similar spacing is also shown.

The addition of metal plates to the top and bottom of the reflector was found to have only a small effect on the v.r.p. These plates provide a convenient mechanical support for the slotted cylinder and also assist in protecting the aerial from the weather.

(3.3) Gain

To compute the aerial gain exactly would have required radiation-pattern measurements in several planes. It was sufficiently accurate, however, to estimate the gain in the following way, using measurements made in only two planes. The gain, relative to a half-wave dipole, which the aerial would have if excited by one slot was taken to be the ratio of the maximum field to the r.m.s. field of Fig. 4. To this was added the gain of four collinear half-wave dipoles having the same spacing as the slots. This is justifiable because the v.r.p. of the complete aerial, at all bearings, would be sufficiently similar in shape to that of four collinear half-wave dipoles in free space. This gives the intrinsic gain of the aerial; the net gain is slightly less on account of loss in the coaxial feeder system inside the tube. The net gain was deduced as follows:

	dB
Estimated gain of aerial if excited by one slot, relative to a $\lambda/2$ dipole	9·3
Theoretical gain of 4 collinear dipoles spaced $1 \cdot 0\lambda$, relative to a $\lambda/2$ dipole	6·5
Intrinsic gain, relative to a $\lambda/2$ dipole	15·8
Loss in coaxial feeder system	0·7
Net gain, relative to a $\lambda/2$ dipole	15·1

The gain of the aerial made for use at 900 Mc/s was not measured, but a similar estimate of gain made for a later aerial, modified for use at 774 Mc/s, agreed with the measured value to within 0·4 dB.

(4) ADMITTANCE

The parameters available for admittance matching were (*a*) variations in the distance of the slotted cylinder from the apex of the corner reflector, (*b*) changes of slot length and width, (*c*) addition of distributed capacitance between the slot faces, (*d*) addition of cylindrical sleeves to the inner tube of the slotted cylinder on each side of the centre spacer, and (*e*) variation in the length of the series compensating stub.

The presence of the reflector modified the input admittance of the slotted cylinder. With the slot dimensions shown in Fig. 2, a satisfactory admittance characteristic was achieved when the cylinder was mounted with its axis $0 \cdot 6\lambda$ from the apex of the reflector. With this spacing the radiation-pattern requirements

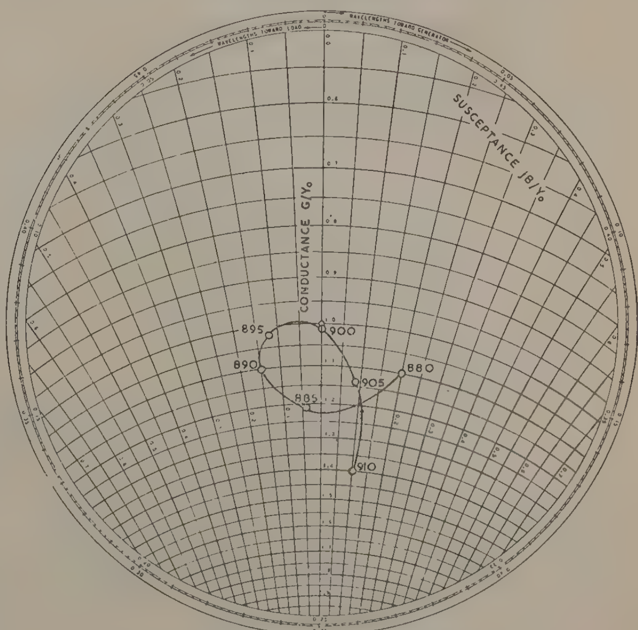


Fig. 6.—Input admittance of the 900 Mc/s aerial.
Figures against the points indicate frequencies in Mc/s.
Admittance normalized to 14·0 millimhos.

were also satisfied. To improve the admittance characteristic the slots were fed via a series-connected stub, adjusted in length to provide the correct reactance compensation.

The resulting input-admittance characteristic is shown in Fig. 6.

(5) OPERATION AT OTHER FREQUENCIES

Minor modifications can be made to the slotted cylinder which enable it to be used, in the same reflector, at other frequencies.

The aerial described was later modified for use at 774 Mc/s by loading the internal feeder system with shunt capacitance in the form of polystyrene sleeves. This reduced the velocity of propagation within the cylinder and re-established a co-phased voltage distribution at the slots. The aerial was matched at 774 Mc/s by adjusting the parameters discussed in Section 4. The net gain was necessarily less than at 900 Mc/s owing to the reduced aperture; the measured value was 13·4 dB.

(6) CONCLUSIONS

The results show that the use of an array of slots as the exciting element for a corner reflector leads to efficient use of the reflector aperture. An aerial made to this design functioned well in service; its performance was unaffected by climatic conditions.

It has been shown that simple modification to the slot loading and feeder system enables the aerial to be used at any frequency within a frequency range of 1·16 : 1. If this range were extended, an upper limit of 1·5 : 1 would be set by the dielectric constant of the polystyrene which is used to reduce the phase velocity within the cylinder.

(7) ACKNOWLEDGMENTS

The authors wish to thank Mr. G. D. Monteath and Mr. P. Knight, who gave much helpful advice as the work proceeded; and Mr. G. B. Collins, who carried out the modifications when it was required to operate the aerial at 774 Mc/s.

The authors are also indebted to the Director of Engineering of the British Broadcasting Corporation for his permission to publish the paper.

(8) REFERENCES

(1) Moullin, E. B.: 'Radio Aerials' (Clarendon Press, 1949).

(2) Beam, R. E., and Ross, H. D.: 'Impedance and Radiation Characteristics of Slotted-Cylinder Antennas', *Proceedings of the National Electronics Conference*, 1949, **5**, p. 172.

21.376.22

The Institution of Electrical Engineers
Paper No. 3295 E
Sept. 1960

A CORRELATOR EMPLOYING HALL MULTIPLIERS APPLIED TO THE ANALYSIS OF VOCODER CONTROL SIGNALS

By Professor A. R. BILLINGS, B.Sc., Ph.D., Associate Member, and D. J. LLOYD, B.Sc.(Eng.), Ph.D., Graduate.

(*The paper was first received* 16*th April, and in revised form* 24*th June*, 1959.)

SUMMARY

After an initial discussion of short-term correlation functions the design and performance of an auto- and cross-correlator using Hall-effect multipliers are described. It is shown that these devices are well suited to act as the multiplying elements and as modulators in associated circuits, provided that one of the time functions to be multiplied has a narrow bandwidth. Examples of measured auto- and cross-correlation functions are given for sine waves and noise and for the control signals of a vocoder, from which interesting conclusions can be drawn about the redundancy of vocoder signals.

(1) INTRODUCTION

In recent years the use of correlation techniques has grown and many correlators of differing design have evolved.[1, 2] Recently, whilst studying speech-analysis systems, it was found desirable to provide a correlator to extract useful information from vocoder control signals. To this end, a correlator has been built which is suitable for analysing signals confined to the frequency band 0–100 c/s. The correlator differs from its predecessors in using the direct multiplying action of a Hall plate, both in the multiplication unit and in modulators in subsidiary circuits. This has led to a very simple circuit design.

(2) SHORT-TERM AUTO- AND CROSS-CORRELATION

The general correlation function $\psi_{ab}(\tau)$ for two time functions $g_a(t)$ and $g_b(t)$ is defined as

$$\psi_{ab}(\tau) = \lim_{T\to\infty} \frac{1}{T}\int_0^T g_a(t)g_b(t-\tau)dt \quad . \quad . \quad . \quad (1)$$

When $g_a(t)$ and $g_b(t)$ are different, the resultant function is termed the *cross-correlation function*, and when they are the same it is termed the *auto-correlation function*.

The functions defined by eqn. (1) are very useful mathematical tools when solving circuit problems, but when making practical correlation measurements some modified definition is required, since T must then be finite. There are two practical possibilities: either a continuous analysis can be made of live signals; or an analysis can be made upon signal samples of finite length repeated periodically. These possibilities lead to two different definitions for short-term correlation functions.

When a continuous analysis is performed, the short-term correlation function λ_{ab} is a function of both τ and t. With a perfect integrator integrating over the previous T_1 seconds, the short-term correlation function is

$$\lambda_{ab}(\tau, t) = \frac{1}{T_1}\int_{t-T_1}^{t} g_a(t')g_b(t'-\tau)dt' \quad . \quad . \quad . \quad (2)$$

In practice, perfect integrators are never available, and imperfect integrators consisting of inductance-resistance or resistance-capacitance networks are commonly employed. When the input to such an integrator with time-constant T_1 is $g_a(t')g_b(t'-\tau)$, the output is

$$\lambda'_{ab}(\tau, t) = \frac{1}{T_1}\int_{-\infty}^{t} g_a(t')g_b(t'-\tau)\exp\left(\frac{t'-t}{T_1}\right)dt' \quad . \quad (3)$$

This will be termed the *weighted short-term correlation function*.

When a repeated analysis is made, the correlation function then applies to periodic functions $g'_a(t)$ and $g'_b(t)$, each having period T_2. With a perfect integrator, integrating over a period $T_1 \gg T_2$, the short-term correlation function is

$$\phi_{ab}(\tau) = \frac{1}{T_2}\int_0^{T_2} g'_a(t')g'_b(t'-\tau)dt' \quad . \quad . \quad . \quad (4)$$

provided that $\tau \ll T_2$ and end-effects can be ignored. If an imperfect integrator of the type already mentioned is used and the time-constant T_1 is not very much greater than T_2, ϕ_{ab} is a periodic function of t with period T_2. Within the period between $t = 0$ and T_2, the modified short-term correlation function can be defined as:

$$\phi'_{ab}(\tau, t) = \frac{1}{T_1}\left[\int_0^t g'_a(t')g'_b(t'-\tau)\exp\left(\frac{t'-t}{T_1}\right)dt' + \int_0^{T_2} g'_a(t')g'_b(t'-\tau)\exp\left(\frac{t'-t}{T_1}\right)\sum_{n=1}^{\infty}\exp\left(\frac{-nT_2}{T_1}\right)dt'\right]. \quad (5)$$

which simplifies to

$$\phi'_{ab}(\tau, t) = \frac{\exp(-t/T_1)}{T_1[1-\exp(-T_2/T_1)]} \times \left[\int_0^t g'_a(t')g'_b(t'-\tau)\exp\left(\frac{t'}{T_1}\right)dt' + \int_t^{T_2} g'_a(t')g'_b(t'-\tau)\exp\left(\frac{t'-T_2}{T_1}\right)dt'\right] \quad (6)$$

As has already been pointed out $\phi'_{ab}(\tau, t)$ is periodic in t. To distinguish it from $\lambda'(\tau, t)$, the function $\phi'(\tau, t)$ will be referred to as the *periodic weighted correlation function*.

It can be seen from eqn. (6) that when $T_1 \gg T_2$ the periodic component is small. The correlator to be described later performs an integration of the form shown by eqn. (6), and $\phi'_{ab}(\tau, t)$ is displayed as an instrument reading. It is convenient when measuring $\phi'_{ab}(\tau, t)$ over a range of τ to keep t fixed, and the simplest value to choose is $t = T_2$. Eqn. (6) then simplifies to

$$\phi'_{ab}(\tau, T_2) = \frac{\exp(-T_2/T_1)}{T_1[1-\exp(-T_2/T_1)]}\int_0^{T_2} g'_a(t')g'_b(t'-\tau)\exp\left(\frac{t'}{T_1}\right)dt'$$

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.

Dr. Billings and Dr. Lloyd were formerly in the Department of Electrical Engineering, University of Bristol.
Dr. Billings is now Professor of Electrical Engineering, University of Western Australia, and Dr. Lloyd is with the National Research Council, Ottawa.

(2.1) Requirements of Practical Correlator

From eqn. (1) it can be seen that a practical correlator has three functions to perform: it must store and delay one input signal; it must multiply this delayed signal by another undelayed signal; and it must integrate the product. The effect upon correlation measurements of finite integration time and imperfect integrators has already been discussed, and it can be seen that, using a simple integrator, a meaningful weighted correlation function can be obtained, and therefore integration presents no difficulties. The delay of one signal with respect to the other is most easily obtained by the use of magnetic-tape storage and is a standard technique. The main difficulty in designing a correlator is in the multiplication prior to integration.

(2.2) Choice of Multiplier

There are many forms of multiplier available, most of which are complicated and operate in an indirect manner.[1–3] In recent years a well-known device which performs a direct multiplication has become a practical possibility as a multiplier element. This is the Hall plate, which gives an output proportional to the product of a flux density and a current.[4] With earlier Hall-plate materials the power available from a Hall multiplier was very small, but with the advent of new compounds such as indium antimonide it is now possible to get high conversion efficiencies, and a Hall multiplier can supply an integrator directly without intervening amplifiers.

The obvious advantage of the Hall multiplier is its extreme simplicity, but it has countervailing disadvantages. The first of these is that magneto-resistance, feedback within the device and technological imperfections lead to some errors in multiplication. These have been discussed in detail elsewhere,[5] and it is sufficient at the moment to say that the overall error can be kept to less than 2%. Another disadvantage is that one of the variables must be represented by a flux density. This means that, if one of the input signals contains frequencies from zero upwards, there is a severe limitation imposed on the permissible upper frequency, since stored energy must increase directly with frequency. In the present context, when designing a correlator to operate on vocoder control signals, this latter disadvantage is not serious since the vocoder control signals do not contain components at frequencies above 25 c/s.

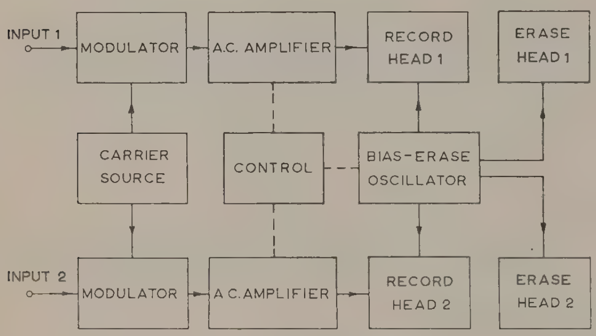


Fig. 1.—Recording system.

(3) CORRELATOR DESIGN

The correlator which has been constructed to analyse vocoder control signals consists of three parts, namely the input circuits, the storage and delay unit and the output circuits. Because of the low frequencies involved, direct recording of the vocoder control signals on the tape is not feasible, and an a.m. system is used with a carrier frequency of 2 kc/s. Twin channels are employed for all stages preceding the multiplier. Diagrams of the correlator circuit arrangements up to the recording heads and subsequent to the replay heads are shown in Figs. 1 and 2.

In the input unit, the signals to be analysed first modulate a 2 kc/s carrier, and Hall modulators are used for this operation. The resulting signals are then amplified and passed to the recording heads in the storage and delay unit. The input unit also contains a bias-erase oscillator for the tape system, and a recording control circuit which switches the input circuit in the correct sequence.

The storage and delay unit contains a closed loop of standard ¼ in tape and two tracks are in simultaneous use, one for each channel. Three heads—erase, record and replay—are used for each track, making a total of six. Of these one of the replay heads is movable along the tape and provides a variable delay.

In the output unit, signals from the replay heads in the tape unit are amplified and demodulated in separate channels. The resulting low-frequency signals are fed, through d.c. amplifiers, to the Hall multiplier, after which the integration of the product is combined with its display in a heavily damped galvanometer.

When measuring periodic weighted correlation functions, the waveform to be analysed is recorded on the tape and the value of τ is set by moving the adjustable replay head. The value recorded for the correlation function is then the reading of the galvanometer at the end of a complete tape traverse, i.e. $\phi'(\tau, T_2)$.

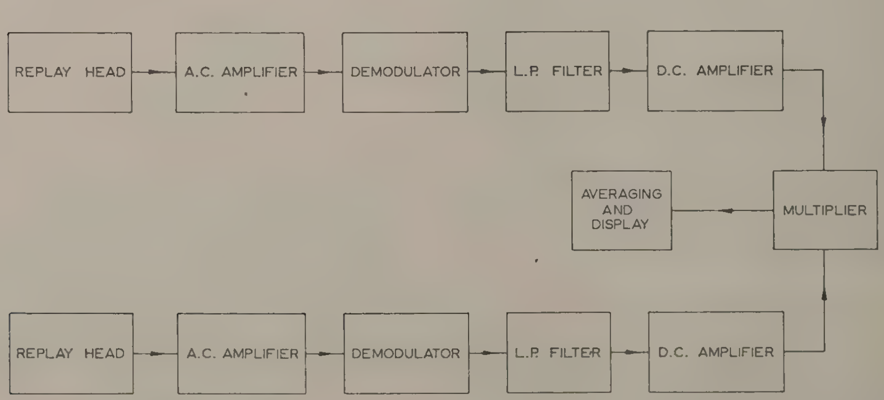


Fig. 2.—Replay system.

Special care has been taken to reduce errors in the multiplier. o keep non-linearity due to magneto-resistance to a minimum, e Hall plate is fed from a high-current amplifier of relatively gh output impedance[6] and a similar mismatch is introduced tween the Hall-plate output and the integrating galvanometer. ne Hall plate is mounted in the air-gap of a ferrite pot-core assembly,[7] and magnetic coupling between the two inputs and the output is minimized by using a lead arrangement as shown in Fig. 3. To remove the unbalance output in the absence of an input to the field, a Kuhrt balancing network[5] is used.

Fig. 3.—Hall-plate lead arrangement.

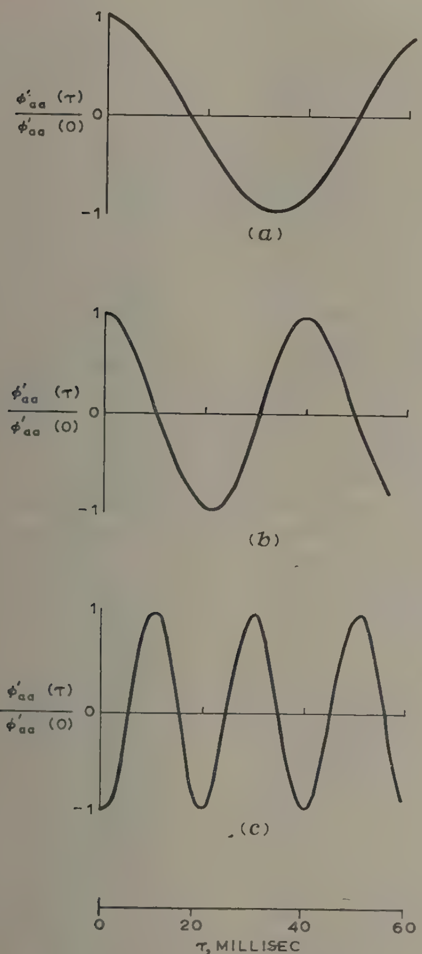


Fig. 4.—Correlograms for sine waves.

(a) Auto-correlogram for 15c/s signal.
(b) Auto-correlogram for 25c/s signal.
(c) Cross-correlogram for 50c/s signal.
$g_a(t) = -g_b(t)$

(4) EXAMPLES OF CORRELATION ANALYSIS

The correlator has been used to measure the periodic weighted auto-correlation function for sine waves, for noise and for sine waves plus noise. The measured correlograms are shown in Figs. 4, 5 and 6 and correspond to what can be predicted by mathematical analysis.[8] They indicate the reliability of the

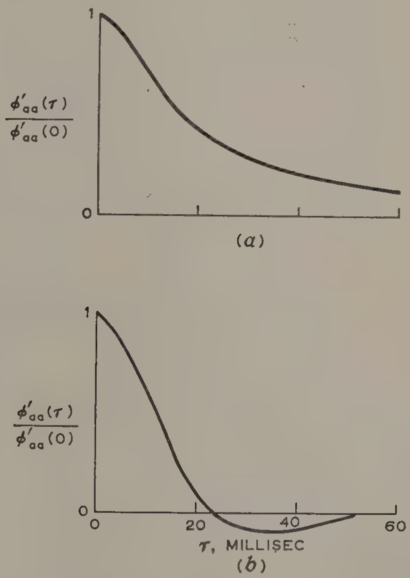


Fig. 5.—Auto-correlograms for envelope detected noise.

(a) Original noise bandwidth, 16c/s.
(b) Original noise bandwidth, 32c/s.

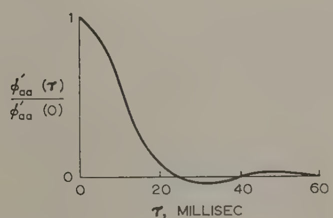


Fig. 6.—Auto-correlogram for envelope detected carrier plus noise.

R.M.S. carrier equal to r.m.s. noise. Original noise bandwidth, 32c/s.

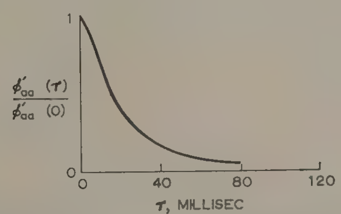


Fig. 7.—Auto-correlogram for vocoder control signal when input is white noise.

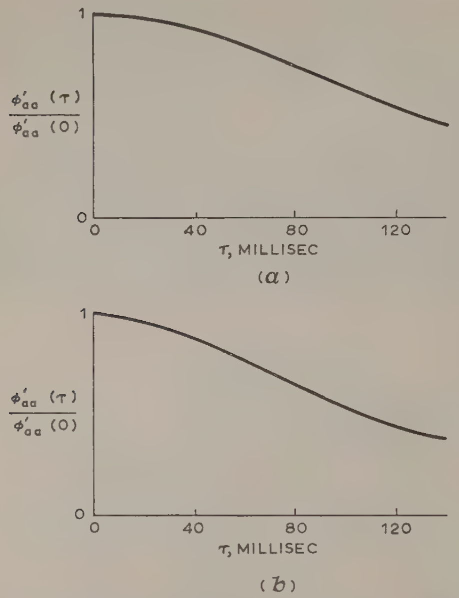


Fig. 8.—Auto-correlograms for vocoder control signals.

(*a*) 'Lah' on channel 5.
(*b*) 'Oi' on channel 7.

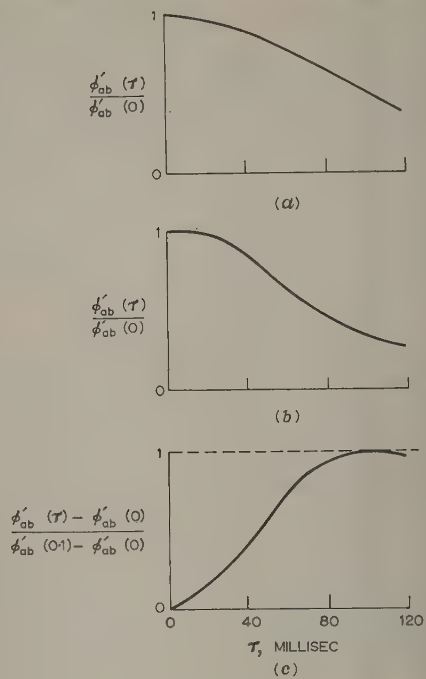


Fig. 9.—Cross-correlograms for vocoder control signals.

(*a*) 'Lah', channels 2 and 5.
(*b*) 'Oi', channels 2 and 7.
(*c*) 'Joe', channels 2 and 10.

correlator. The quantity plotted in these diagrams is the normalized correlation function $[\phi'_{aa}(\tau)/\phi'_{aa}(0)]$, where the time of measurement, T_2, is implied.

Finally, some preliminary measurements of both auto- and cross-correlation functions for vocoder control signals have been made and are shown in Figs. 7, 8 and 9. A more detailed study of such correlograms is now being made, but some conclusions can be drawn from the preliminary measurements. The correlogram in Fig. 7 is for a random-noise input to the vocoder with a pass band of 25 c/s for the derived control signal. It can be seen that the correlograms for speech inputs are much less sharp than that for the noise input. This indicates that, for the sounds analysed, the control-signal power spectrum occupies a bandwidth considerably less than the 25 c/s commonly accepted as necessary. Another conclusion can be drawn from the cross-correlation measurements, namely that there is still considerable redundancy present in vocoder signals. This can, of course, be inferred from the high channel capacity required by vocoder signals, but it is hoped that further study of correlograms may reveal how some of the remaining redundancy can be removed. Of particular interest amongst the correlograms already measured is that of Fig. 9(*c*), which shows a pronounced delayed correlation.

(5) ACKNOWLEDGMENT

The authors would like to express their thanks to Professor G. H. Rawcliffe for the experimental facilities provided at Bristol University.

(6) REFERENCES

(1) LEE, Y. W., CHEATHAM, T. P., and WIESNER, J. B.: 'Applic tion of Correlation Analysis to Detection of Period Signals in Noise', *Proceedings of the Institute of Rad Engineers*, 1950, **37**, p. 1165.

(2) HOLMES, J. N., and DUKES, J. M. C.: 'A Speech-Wavefor Correlator with Magnetic-Tape Delay and Electron Multiplication', *Proceedings I.E.E.*, Paper No. 1639 July, 1954 (**101**, Part III, p. 225).

(3) JONES, J. C., and READSHAW, D.: 'Lace (The Luton Analog Computing Engine), Part 2', *Electronic Engineering*, 195 **29**, p. 380.

(4) ROSS, J. M., and SAKER, E. W.: 'Applications of Indiu Antimonide', *Journal of Electronics*, 1955, **1**, p. 223.

(5) BILLINGS, A. R., and LLOYD, D. J.: 'The Sources of Err in Hall-Effect Multipliers, with particular reference their Use as Modulators', *Proceedings I.E.E.*, Pap No. 2989 E, February, 1960 (**106** B, Suppl. 16, p. 706).

(6) LLOYD, D. J.: 'A Simple Transistor Amplifier for Energizi a Hall Multiplier', *Electronic Engineering*, 1958, **30**, p. 56

(7) LLOYD, D. J.: 'Pot-Core Construction for a Hall Multiplie *Journal of Scientific Instruments*, 1958, **35**, p. 225.

(8) BURGESS, R. E.: 'Observation of Signals in the Presence Noise', *Philosophical Magazine*, 1951, **42**, p. 493.

1.396.97 : 534.84

The Institution of Electrical Engineers
Paper No. 3272 E
Sept. 1960

DIFFUSION OF SOUND IN SMALL ROOMS

By K. E. RANDALL, Graduate, and F. L. WARD, B.Sc., A.Inst.P., Associate Member.

(*The paper was first received* 13*th January, and in revised form* 1*st April,* 1960.)

SUMMARY

The paper describes an investigation of some problems of sound ffusion in rooms, with particular reference to small sound and telesion studios. The experimental work shows that it is possible to easure the degree of diffusion in a room by fairly simple practical chniques. Quantities based on the frequency variation of reverberaon time and double reverberation decay constants are the most omising for use in small broadcasting studios.

It is also shown that uniform distribution of absorption can be as fective as other means of attaining conditions of good diffusion. ectangular diffusers are particularly effective in improving conditions here the distribution of absorption is poor.

(1) INTRODUCTION

(1.1) Significance of the Term 'Diffusion'

Experience shows that a number of undesirable subjective fects in broadcasting studios are associated with an inadequate egree of diffusion. A sound field is defined as being comletely diffuse if it has uniform energy density within the region onsidered and if the directions of propagation at any arbitrarily elected points are wholly random in distribution. Although erfect diffusion is thus identifiable theoretically, there is no enerally accepted method of describing quantitatively a state f diffusion which falls short of perfection. A previous paper as dealt with some of the aspects of this problem.[1]

The importance of a knowledge of the state of diffusion lies the fact that reverberation formulae, e.g. those of Sabine and yring, assume perfect diffusion in a room.[2] Experiments have own that in normal rooms of reasonable size and for wavengths appreciably smaller than the dimensions of the room, e formulae are in general valid. The most important use of ese formulae in broadcasting is to enable the behaviour of absorbing material in a studio to be predicted from its ehaviour in a reverberation room. Since there have been cases here such predictions have proved to be inaccurate, an examinaon of the mechanism of diffusion is important for this reason lone. The work described was carried out under laboratory onditions, but attention has been directed throughout to finding ractical means for investigating studios where time is limited nd apparatus must be easily portable.

The term 'diffuse' refers to the sound field itself but the term iffused' is used to describe a room where, in the frequency and under consideration, the sound field attains a high degree f diffusion. The term 'diffuser' is used to describe a rigid bject possessing a negligible sound-absorption coefficient fixed a flat surface of a room so as to change the shape of the rface.

A room can be considered as a system having a large number f modes of resonance within the audible frequency range. The xcitation of each mode is determined by the position of the ource of sound in the room and its frequency, whilst its damping determined by the absorbing power of the surfaces associated ith it. The absorption taking place in a given surface material will depend, among other factors, on the orientation of the surface with respect to the wavefront.

A pure tone present in a room excites a number of modes. The excitation of a particular mode decreases as the frequency interval between the tone and the resonance frequency of the mode increases; in practice, therefore, the number of modes contributing significantly is limited. The intensity level of sound in a room is at all times the vector sum of the modal responses and the exciting tone.

It is customary in the study of the transient (reverberant) response of a room to observe the effect of switching off a steady tone. It should be noted that before switching off, the frequency of oscillation, i.e. the forced oscillation, is that of the exciting tone, but at the instant of cut-off each of the modes of the room has associated with it an immediate response in the form of a decay of tone at its own modal frequency.

The resultant decay follows an approximately exponential law where there are a large number of equally excited and equally damped modes substantially sharing the sound energy: there are, however, beats between the modes which are energized at different frequencies, resulting in a modulation of the envelope of the sound decay.

It is convenient at this stage to consider the logarithm of the sound pressure level, i.e. to use a decibel scale. When this is done, a sound decay which is exponential with respect to time becomes a straight-line decay, and oscillations in the exponential envelope appear as oscillations about a mean straight line. In the same way changes of exponential decay constant during the decay appear as changes in slope of the straight line; when the changes are gradual the line exhibits a curvature.

Thus, while modulations of the straight-line decay occur as a result of beats between different modes, double-slope or curvature effects may be evident where the component modes do not all have the same damping coefficient. As the number of contributory modes increases, the beats become less evident and curvature effects disappear, because the shape of the decay tends to be controlled by the more lightly damped modes. The decay thus tends to become linear.

It should be noted that a pure straight-line decay may also correspond to the excitation of a single mode. This extreme case cannot occur in a room, although conditions approaching it do occur when the modes are so widely separated in frequency that contributions from adjacent modes are negligible. It is important to distinguish between this type of straight-line decay and that described above.

In the treatment above, we have described the sound field in a room in terms of a system of modes. Using this concept we are led to the conclusion that perfect diffusion is obtained only when the sound energy in the room is carried equally by an infinite number of modes. Although unattainable in practice, this idea is useful as a norm in assessing the degree of diffusion.

(1.2) Mechanism of Diffusion and Steady-State Characteristic

Resonance arises as a result of successive reflections of sound waves from the room surfaces. The frequencies of the modes of resonance are determined by the time delays between reflections

Written contributions on papers published without being read at meetings are vited for consideration with a view to publication.
Mr. Randall and Mr. Ward are in the Engineering Research Department, British roadcasting Corporation.

and hence by the geometry of the room. Where the absorbing materials are purely resistive and of negligible depth, the modal frequencies can be only slightly affected by the distribution of absorption over the surfaces. The sound energy associated with each mode, however, will be governed by the absorption coefficients of the surfaces involved in its development. In general, it can be expected that the most uniform distribution of energy between the modes (implying the most uniform distribution through the room) will occur when the absorption coefficient is the same over all the surfaces, because otherwise energy will be concentrated into particular modes. The term 'intrinsic' is used, for the purposes of the paper, to describe the degree of diffusion attained under these conditions.

A lower degree of diffusion for a given frequency region will naturally be obtained if the distribution of absorption is poor; an appreciably higher degree can be achieved only by changing the shape of the room in some way. A change in the ratio of dimensions may improve the diffusion in a particular frequency region, although possibly at the expense of another part of the audio spectrum. Irregularities introduced into the walls, such as rectangular or other projections, improve the degree of diffusion by separating the modes into a larger number of components.

The above considerations may be summarized in the statement that in a given room, where the walls are plane and absorption is not perfectly distributed, the degree of diffusion can be improved up to a maximum intrinsic value by improving the uniformity of the average absorption between the surfaces. Irregularities in the walls, such as are provided by artificial diffusers, are the only means by which the degree of diffusion may be increased beyond this value.

This simple exposition ignores effects due to diffraction occurring, for example, at the edges of patches of absorber and the edges of diffusers. Such effects imply a change of shape of the wavefront and will increase the degree of diffusion to a small extent because the direction of propagation of part of the energy is altered.

The importance of this factor cannot be accurately assessed until further work has been carried out. In the work described the possibility of diffraction effects was borne in mind and the experiments were so devised that as far as possible any such effects would not invalidate the conclusions.

A study of the steady-state frequency response of a room is perhaps the most obvious line of attack for finding a measure of the state of the sound field, as the peaks in the response can be related directly to the resonance frequencies of the modes. This has been used in earlier investigations.[1, 3]

When the source of sound is in one corner of the room and the receiver in another, the peaks of the steady-state response correspond to the mode frequencies, provided that the modes are clearly separated, i.e. if the room is small or the frequency low. As the frequency is increased and the modal separation therefore reduced, the picture becomes complicated by the contributions of adjacent modes. If a state of perfect diffusion existed the modes would not be identifiable in any way.

Corresponding to the steady-state frequency response for any one point in a room, there exists a steady-state spatial distribution in the room for any one frequency. This might also be used as a diffusion indicator.

(1.3) Measurements from Fluctuations in Decay Curves

Beats in the envelope of a sound decay when tone has been cut off will tend to disappear as the degree of diffusion increases; an examination of the amplitude of these beats might therefore lead to some kind of diffusion index. Analysis of this type has indeed been used in the assessment of concert halls, and a laboratory development of the method was the starting-po for the present investigation. Although results from stud and concert halls and some simple laboratory experiments h shown promise, it was clear that a very large number of dec needed to be analysed before any reliable, repeatable resu could be obtained. For this reason an attempt was made develop an automatic method[4] for measuring a quantity defined as the modulus of the mean deviation in decibels of fluctuating decay curve from the best-fit straight line.

The method showing most promise involved measuring not factor D itself but a related quantity. For this purpose modification of the equipment built for the measurement steady-state level irregularity was used.

This equipment was based on a high-speed level recorder a measured the average amplitude of oscillations in the straig line decay. Readings of this quantity, called for convenie the decay irregularity, were unfortunately very critically dep dent on the adjustments of the apparatus. Several room co ditions were investigated and it was found that, althou differences could be detected by very careful control of conditions of the experiment, especially reverberation time, general these differences were of the same order as the over errors of measurement.

The average values of the decay irregularity for three con tions are given in Table 1. These conditions were chosen represent the maximum possible range of diffusion.

Table 1

RESULTS OF MEASUREMENT OF DECAY IRREGULARITY

Condition	*Reverberation time*	*Decay irregula*
Room with rectangular diffusers ..	1·0	15·4
Room empty	1·8	16·9
Room with absorption concentrated on one wall	0·9	19·0

The differences between these conditions are not very lar although they are certainly significant. The highest dec irregularity is obtained with a condition where a low degree diffusion must be expected and the lowest value where st have been taken to increase diffusion. In particular, the l value is lower than that for the empty room, which, hav uniformly distributed absorption, provides an example of a ro with an intrinsic degree of diffusion.

In practice, the technique just described appeared to be far t insensitive for practical purposes and work on it was theref discontinued. A new approach was made to the problem finding a practical diffusion index. It was decided to set conditions representing practical extremes of absorption distrib tion and diffusion, and then to determine what correspondi changes had taken place in the sound field by examining all quantities which seemed likely to be affected by the degree diffusion, other than steady-state and decay irregularity alrea discussed.

Following from the general concept of diffusion, seve properties of a perfectly diffuse sound field can be assumed:

The frequency and spatial irregularities obtained from stea state measurements will be negligible.

There will be negligible modulation of the decay characteristics the form of beats over a wide range of frequencies.

Decays will be perfectly exponential, i.e. they will be represen by simple straight lines on a logarithmic scale.

The form of the decay will be independent of the measur position in the room.

The character of the decay will not be critically dependent frequency.

The character of the decay will be independent of the directio component of velocity used for measurement.

The first two of these characteristics having already been tudied, attention was concentrated on the remainder of these roperties of a perfectly diffuse sound field. The first series of xperiments was restricted to a study of the average slope of the ecay, measured as reverberation time.

(2) FIRST SERIES OF MEASUREMENTS USING REVERBERATION-TIME STATISTICS

(2.1) Methods of Measurement and Analysis

All tests were carried out in a tiled room approximating to a 0ft cube. The prefix A will be used to refer to measurements arried out in this room.

Pulses of tone were produced by a loudspeaker placed in one orner of the room. An omnidirectional microphone was con- ected through a microphone amplifier to a logarithmic amplifier nd thence to an oscillograph on which the sound decays could e photographed. Two types of record were produced, first vith warble and secondly with pure tone pulses.

With warble tone pulses, the deviation was $\pm 10\%$ and the varbling frequency 6c/s about each of seven frequencies from 00c/s to 8kc/s. Fifteen microphone positions were used, listributed throughout the room and including three corner ositions. The total number of recorded decays was thus 105 or each condition.

With pure tone pulses, starting at each of four frequencies 700c/s, 1kc/s, 1·4kc/s and 2kc/s), 26 pulses at 2c/s intervals vere recorded using a microphone placed in the centre of the oom. There were thus 104 decays for each condition. The hoice of 26 pulses covering a band of 50c/s was intended to rovide a sufficiently detailed exploration in the region of each requency while not extending so far as to involve any changes f mean absorption between the ends of the frequency band.

The average slope of each of the decays was measured from he record and converted to reverberation time. In order to educe the effect of variations due to individual interpretation of he slope of a decay by a particular observer, each decay was neasured by two observers, the two readings being averaged in ubsequent computation.

Two quantities were computed:

(*a*) From the first record, a quantity P representing the ariation of slope with microphone position. This was the tandard deviation of the 15 individual decays at each frequency xpressed as a percentage of their mean slope.

(*b*) From the second record, a quantity F representing the ariation of slope with small arbitrary increments of frequency. This was the standard deviation of the 26 decays in each band, xpressed as a percentage of their mean.

(2.2) Experiments on the Effect of Distribution of Absorbers

The absorber used for all the experiments was a commercial roduct which gives efficient absorption from 500c/s upwards. The experimental conditions are shown in Fig. 1. In condition 1, 65% of one wall was covered with absorber, the other urfaces being untreated; in condition A2, the same area of bsorber was divided into four sections and mounted on four f the six room surfaces; and in condition A3, one of the four ections was removed, leaving one section on each of three walls nutually at right angles.

Condition A1 was intended to represent a condition of bad istribution of absorption, the distribution being much more niform in the other conditions. The reverberation-time plots re shown in Fig. 2, where it will be seen that, although the same mount of absorption has been used in the second condition, ne reverberation time is appreciably shorter, indicating a greater fficiency of absorption. It was thought possible that the

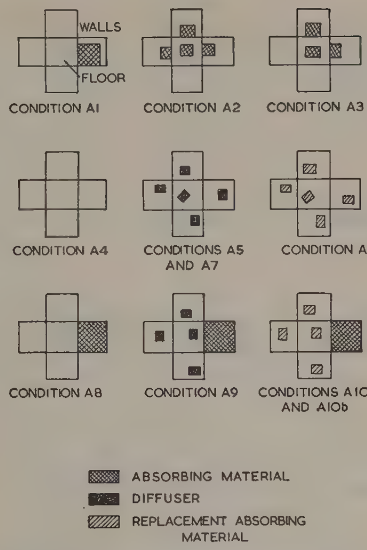


Fig. 1.—Diagrams of room showing disposition of absorbing material and diffusers in various experimental conditions.

Diagrams merely indicate various room conditions. They refer directly to room A, but conditions B8, B9 and B10 in room B were similar to A8, A9 and A10.

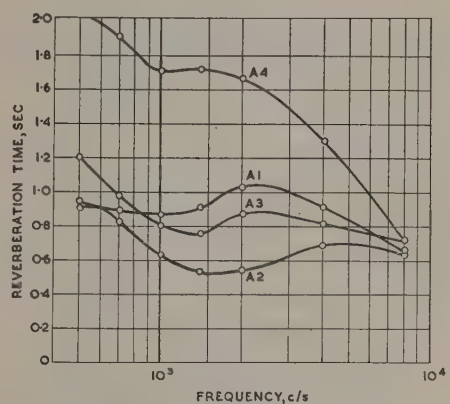


Fig. 2.—Reverberation time of room in experimental conditions A1–A4.

absolute value of the reverberation time might influence the results. It was for this reason that condition A3 was introduced in an attempt to reproduce the reverberation time of condition A1 while substantially maintaining the degree of distribution of condition A2. For comparison another condition, A4, where there was no added absorption, was also investigated. It may be assumed that in this condition the small amount of absorption remaining was perfectly uniformly distributed.

(2.3) Results of Experiments with Absorbers

The values of P and F for the four conditions are shown in Table 2 and plotted in Figs. 3 and 4. The values of P for condition A1, where the absorber is badly distributed, are generally appreciably greater than those for the other three conditions. The result of distributing the absorber (condition A2) is to reduce the value by about 40% throughout the frequency range. In condition A3, where one of the portions of absorber has been

Table 2

Mean Values of P and F with Various Room Conditions

Room condition	P %	F %
A1	11·7	16·8
A2	7·9	—
A3	8·2	10·4
A4	5·3	1·4
A5 (mean of two)	5·7	8·4
A6	6·9	9·9
A7	6·3	8·9
Mean of A5 and A7	6·0	—

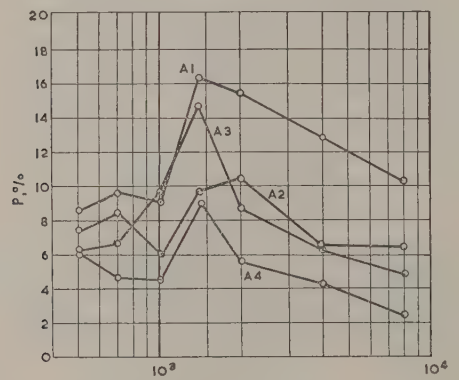


Fig. 3.—Positional variation of reverberation time for experimental conditions A1–A4.

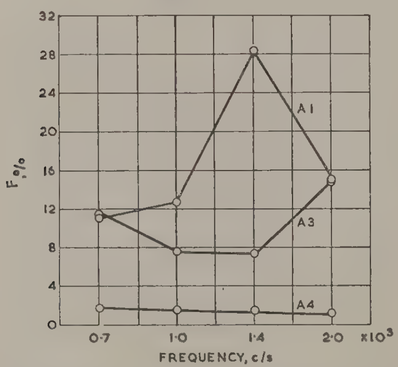


Fig. 4.—Frequency variation of reverberation time, F, for experimental conditions A1, A3 and A4.

removed, the values are of the same order as those from condition A2 except at 1 kc/s and 1·4 kc/s, where the value returns to that obtained in condition A1. The lowest values are given consistently by condition A4 where the room contained no additional absorption. This will be referred to in the paper as the 'empty' state.

It thus appears that the quantity P is in general smaller where the absorption is most uniformly distributed, assuming that the absorption in the empty room has the most nearly uniform distribution.

The values of F for condition A1 are very much greater than those for condition A2 at 1·4 kc/s and 1 kc/s, while the empty condition A4 provides very low values indeed. F therefore appears to be reduced when the absorption is most uniformly distributed.

(2.4) Experiment on the Effect of Diffusers

The second experiment was intended to investigate the effe of introducing rectangular diffusers into the otherwise emp room. In condition A5, 12 diffusers were arranged on five the surfaces of the tiled room. These diffusers were made stout timber finished with a high-gloss paint, and were of thr sizes, the largest being $1\frac{1}{2} \times 2 \times 3$ ft.

The reverberation time in this condition is shown in Fig.

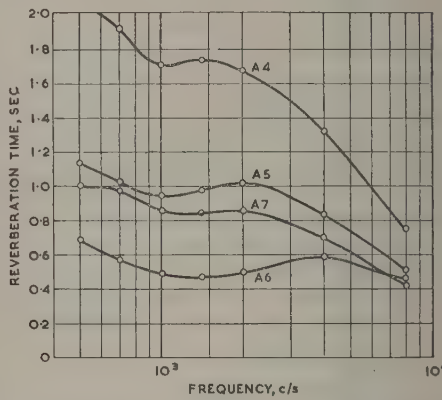


Fig. 5.—Reverberation time of room in experimental conditions A4–A7.

It will be seen that throughout the range it is of the order 50–60% of that of the empty room. It is clear, therefore, th the diffusers introduce additional absorption, and so for cor parison a further condition, A6, was used where the diffuse were replaced by a similar area of absorber.

P and F are shown in Figs. 6(a) and (b) plotted against fr quency, and Table 2 shows the mean values. The values of

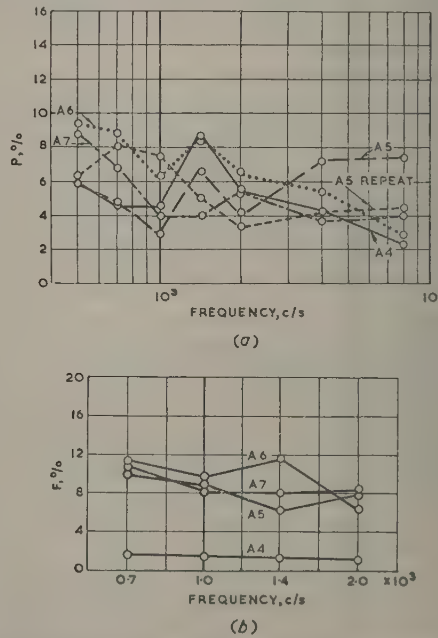


Fig. 6.—Variation of reverberation time for experimental conditio A4–A7.

(a) Positional variation, P.
(b) Frequency variation, F.

›r condition A5, shown in Fig. 6(*a*), are somewhat lower below kc/s than those for the empty room, condition A4, but rise › higher values at 4 kc/s. It was thought possible that this night be associated with a corner reflector effect, the right ngles between the sides of a diffuser and the wall behaving as orner reflectors for sounds of which the wavelength was small ompared with the dimensions of the side. The hypothesis here ; that since a corner reflector returns sound along its incident ath, it tends to maintain the existing order, rather than reating more disorder, and hence to nullify the effects of the iffuser.

In condition A7, therefore, the sides of the diffusers were reated with a thin glass-wool felt which had a high absorption oefficient only at high frequencies. The intention was to reduce he corner reflections by absorbing sound of short wavelength.

(2.5) Results of Experiments with Diffusers

As can be seen from Fig. 6(*a*), the values of *P* for condition \7 are lower than those for condition A5 except below 1 kc/s. ubsequently, condition A5 was repeated with a slightly different rrangement of diffusers. The results are shown also in Fig. 6(*a*). The rise at high frequencies observed in the first neasurement is not seen in this case; in fact, the values are argely the same as those for condition A7 though higher below kc/s.

The variations between these three conditions with diffusers resent suggest that individual values should be treated with eserve, the general trend only being accepted. The mean for onditions A5 and A7 is shown in Table 2. Referring to Table 2 nd Fig. 6(*b*), where the mean values of *F* are shown, it can be een that the values for conditions A5, A6 and A7 are substantially the same, being similar to those for condition A3. These should be compared with condition A4.

(2.6) Further Experiments with Diffusers

In the previous experiments, diffusers as judged in terms of the arameters *P* and *F* did not improve conditions in an empty oom; here a complication arises in that the diffusers themselves ecessarily introduce absorption. In the following experiments he extent to which diffusers could improve an obviously bad ondition was investigated.

One wall of the same room was entirely covered with absorbers condition A8), thus producing a very high degree of spatial symmetry in the distribution of absorption in the room. In ondition A9, 12 diffusers were arranged on the other surfaces f the room, and in condition A10, the diffusers were replaced y patches of absorber. The total absorption contributed by hese patches over the frequency range was similar to that of he diffusers; this was verified by a series of special measurements. As it was found very difficult to simulate the diffuser bsorption precisely, two variations of this condition were used. n condition A10*a*, the amount of absorption was too great and n condition A10*b*, too little. The reverberation-time characteristics measured in the room in these conditions are shown n Fig. 7.

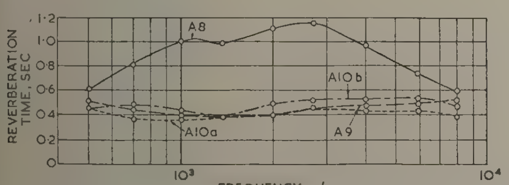


Fig. 7.—Reverberation time of room in experimental conditions A8, A9 and A10.

The results in terms of the parameters *P* and *F* are shown in Table 3 and Figs. 8(*a*) and (*b*). The highest values of *P* are in general evident in condition A8, as was to be expected. The reduction apparent when diffusers are introduced (condition A9) occurs also when the diffusers are replaced by the larger amount

Table 3

Effect of Diffusion on Mean Values of *P* and *F* in Room with Asymmetrical Distribution of Absorption

Room condition	*P* %	*F* %
A8	13·2	15·6
A9	6·6	9·9
A10 (mean of *a* and *b*)	9·0	11·6

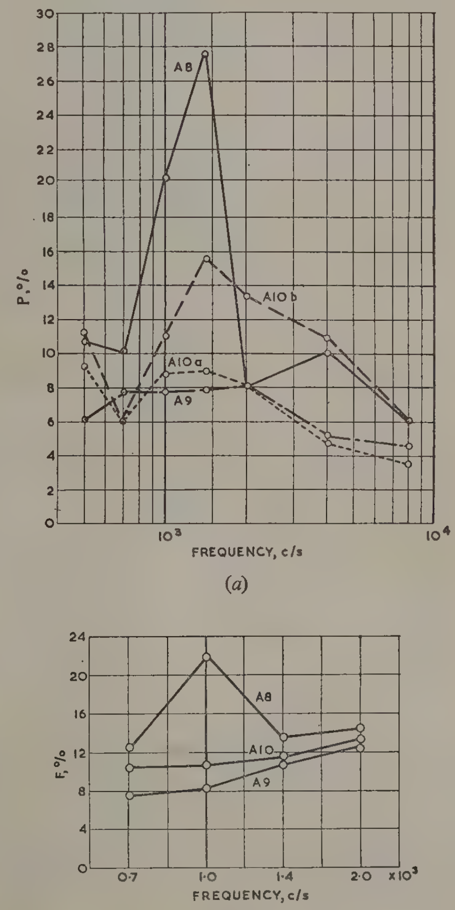


Fig. 8.—Variation of reverberation time for experimental **conditions** A8, A9 and A10.

(*a*) Positional variation, *P*.
(*b*) Frequency variation, *F*.

of absorption; in condition A10*b*, however—that of less absorption—there was a substantial increase in *P* above 1 kc/s compared with condition A9.

The absorption equivalent to that of the diffusers undoubtedly lies between conditions A10*a* and A10*b*. Interpolating values of *P* for these conditions shows that the diffusers have an effect in excess of that to be expected from their absorption alone.

Referring to the values of F observed in the experiments, it is clear that, although there is a substantial reduction with the introduction of diffusers, the equivalent absorbers have nearly as great an effect.

(3) STATISTICS OF REVERBERATION-TIME MEASUREMENTS IN ROOM WITH NON-PARALLEL WALLS

(3.1) Experimental Conditions

In the foregoing measurements certain general trends are perceptible, although there is considerable variation in the results for any particular condition. These variations could be associated with the exact placing of the absorbers or diffusers, and it was therefore thought desirable to repeat the measurements with different arrangements. Repeat experiments were in fact carried out in another completely tiled room, having a volume of 2000 ft^3 and non-parallel walls. Throughout the paper conditions in this non-parallel walled room are prefixed B.

The measurement technique was changed in one respect, in that, instead of reading the decay times directly by means of a logarithmic amplifier and oscillograph, a high-speed logarithmic level recorder was used, giving a paper record. About 100 ft^2 of absorber was mounted on one wall, leaving about 30% of the wall uncovered. This was intended to reproduce condition A8 used in the smaller room and is referred to as condition B8; similarly the diffusers and absorbers were separately introduced to produce conditions B9 and B10. Later these three conditions were repeated with the absorbers and diffusers rearranged.

(3.2) Results

The results given in Figs. 9, 10 and 11, and in Table 4, show a tendency for P to be lower throughout in room B than in the rectangular room. The difference between the three conditions is less clear. Condition B9, with diffusers, is characterized by the lowest values in the high-frequency region for the first configuration and in the middle-frequency region for the second. Although the difference is slight, this condition corresponds to the most consistently low values for the two configurations.

Condition B9 also corresponds to consistently lower values of F than those observed in either of the two other conditions.

Table 4

MEAN VALUES OF P AND F MEASURED IN A NON-PARALLEL-WALLED ROOM

Room condition	P %	F %
B8	8·3	16·4
B9	7·0	9·6
B10	7·1	11·7
B8*	6·7	16·5
B9*	6·3	9·9
B10*	7·5	13·9

* Absorbers and diffusers rearranged.

(4) EXPERIMENTS WITH DIRECTIONAL MICROPHONES

(4.1) Pressure-Gradient Microphone

An investigation of the directional variations in reverberation time with different conditions in a room was carried out with a ribbon microphone. Such a microphone responds to all components of velocity normal to one plane and has a figure-of-eight polar diagram; considered as a directional microphone, therefore, it has a very broad acceptance angle but has the advantage of a complete null in one plane for all frequencies. Conditions A4, A5 and A6 were investigated, both warble tone and pure tone

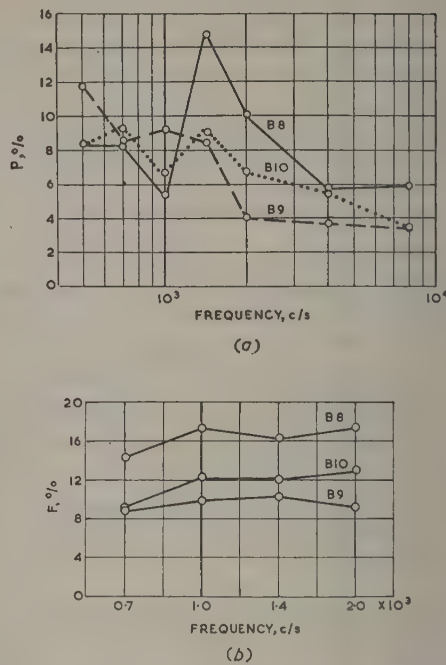


Fig. 9.—Variation of reverberation time for conditions B8, B9 and B10.

(a) Positional variation, P.
(b) Frequency variation, F.

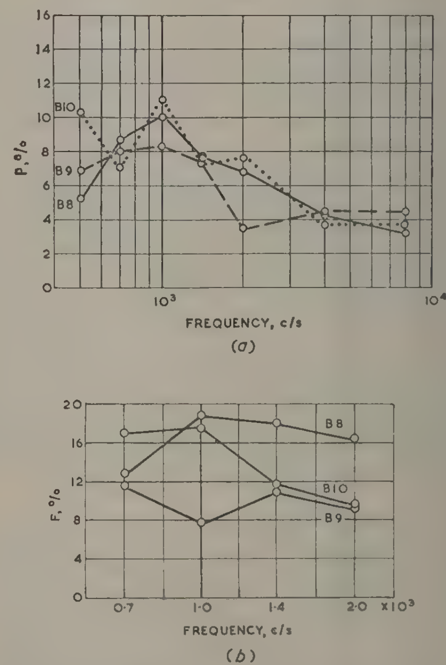
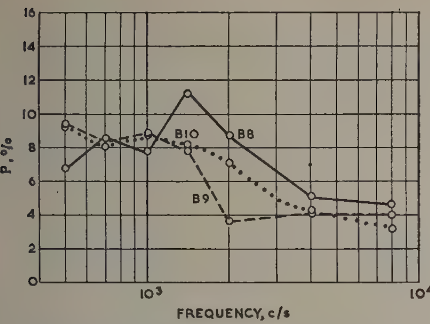


Fig. 10.—Repeat measurements for conditions B8, B9 and B10.

(a) Positional variation, P.
(b) Frequency variation, F.



g. 11.—Positional variation of reverberation time, P, for conditions B8, B9 and B10 (mean of two determinations).

ing used, at frequencies over the range from 175 c/s to 8 kc/s. ith pure tone, two observers made five measurements at ervals of 2 c/s at each frequency, the mean of the ten readings ing taken.

In general any differences with orientation were very slight d the investigation in this respect was not pursued further.

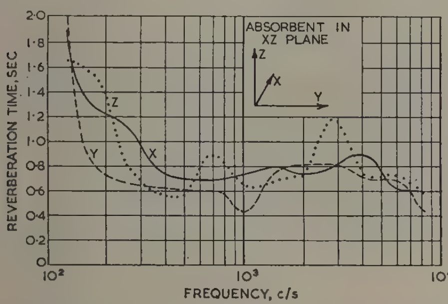


Fig. 12.—Reverberation time measured by pressure-gradient microphone in three directions (experimental condition A8).

Fig. 12 is plotted the result for condition A8 (absorber was ounted on one wall only); this condition, as expected, exhibits e greatest difference between the planes of measurement.

(4.2) Highly Directional Microphone

Subsequently the measurements were repeated using a highly rectional microphone, which unlike a ribbon microphone has unidirectional polar diagram and also a much narrower rward lobe. The microphone consists of a slotted tube, one etre long, containing a graded resistive material, mounted in ont of the diaphragm of a pressure microphone. For sound wavelength less than one metre, partial cancellation takes ace along the length of the tube except where the tube lies ong the direction of propagation of the wavefront. It presents gligible obstruction in the room, but its length is large compared th the wavelength of the standing waves and represents 30% the smallest dimension of the room. The signal obtained om it thus does not represent the sound level at a point but is function of the spatial variation along the length of the icrophone.

The results for conditions A8 and A9 are shown in Figs. 13(a) d (b). As only one microphone position was used, the mean two separate determinations is plotted. In general, for contion A8 there was a more pronounced difference between the sults obtained for different orientations than with the ribbon

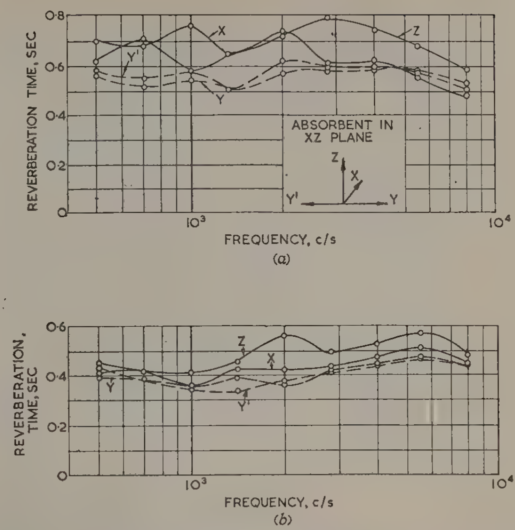


Fig. 13.—Reverberation time using highly directional microphone.
(a) Experimental condition A8.
(b) Experimental condition A9.

microphone. For condition A9 where rectangular diffusers had been added, this difference was appreciably reduced.

(5) MEASUREMENTS OF THE CURVATURE OF INDIVIDUAL DECAYS

An investigation of the curvature of individual decays was confined to a study of the general shape, no attention being paid to the small irregularities. Any change in the exponential index of the decay will appear on the logarithmic scale used for these measurements as a change of slope, which can take one of several forms. The decay may exhibit a single pronounced discontinuity in the slope, several distinct slopes, or even a continuous curve. The shape is usually concave, the slope decreasing with time, but may on occasions be convex. It was arbitrarily decided to treat all decays as having a double slope, and an attempt was made to fit straight lines to the top and bottom halves of the decay. The ratio between the two slopes thus obtained represents the simplest way to express the overall linearity of the curve. The records previously used to determine the quantity F were re-examined and the slope of the upper part of each decay was measured by aligning the reference line of the protractor to it while ignoring the lower part. The slope of the lower part was similarly obtained. The division between the two parts of the curve was set at approximately half-way down the decay, which in practice took into account a range of some 25 dB for each part instead of the normal range of 50 dB for the whole decay. This reduction in the level range is the penalty to be paid for obtaining information about the shape of the curve, and the number of measurements must be increased to attain sufficient accuracy in determination of the slopes. Measurements were made by two observers, whose readings were averaged, before deriving the ratio between the two slopes. For convenience the slopes were measured in terms of the corresponding reverberation time. In Fig. 14, the reverberation times for the two parts of the decay are plotted for one condition, together with the mean values of the ratio, S, which is defined as the ratio of the reverberation time of the high-slope portion to that of the low-slope portion, expressed as a percentage.

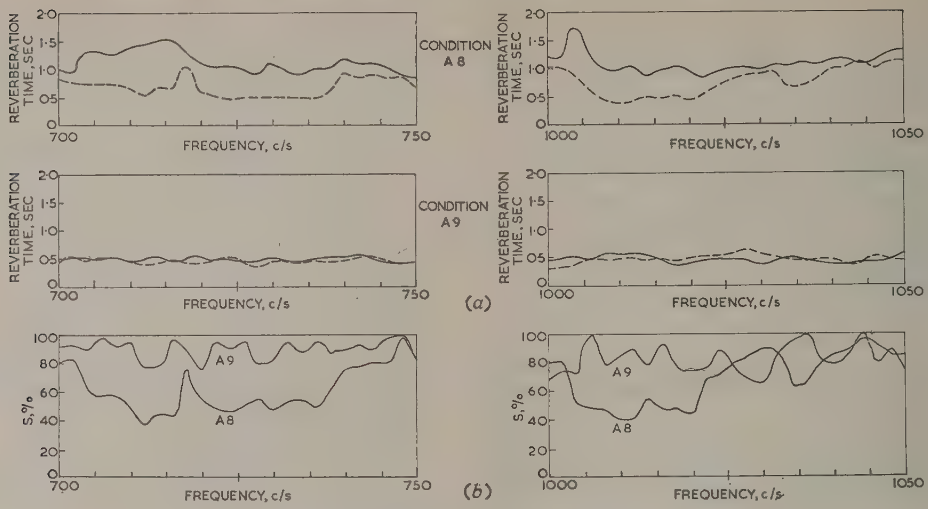


Fig. 14.—Measurement of double slope index, S.

(*a*) Example of measurements of initial and final slope of decay curves.
(*b*) Values of S obtained from the slope measurements.
——— Last part of sound decay.
– – – – First part of sound decay.

The reverberation time of the low-slope portion occurs thus as the denominator irrespective of whether it arises in the first or second part of the decay. It follows that S is never greater than 100% (denoting a straight line), and that a given value of S may refer to either a concave or convex decay. An alternative procedure would be to use the reverberation time from, say, the lower part of the decay always as the denominator. In this case a convex slope would be associated with a figure greater than 100%, while a concave slope would still be associated with a figure less than 100% as above. This procedure would have the disadvantage that, in taking a mean of many individual determinations, departures from the straight line as convex slopes would tend to cancel the effect of departure as concave slopes. The resulting mean would not then be the average of the curvature of the individual decays, which is the quantity required.

In practice, as will be seen from Fig. 14, S varies considerably and not always systematically between decays at 2 c/s intervals, so that it was essential to consider the average of a number of individual readings. Means were computed for each of 26 decays, each group covering a range of 50 c/s starting at one of the four frequencies 700 c/s, 1 kc/s, 1·4 kc/s and 2 kc/s.

Table 5 shows the values obtained in eleven room conditions. Study of the Table shows that there is no sign of a systematic variation with frequency, so that it appears reasonable to consider the mean of readings taken at all four frequencies as representing a value for any particular condition. Standard errors for these overall means are computed from the 104 readings obtained in each case.

Considering first conditions A8, A9 and A10, the lowest value of S corresponding to the highest average curvature was 67%. This occurred in condition A8 in which the only treatment of the room was one wall entirely covered with absorber. When rectangular diffusers were introduced (condition A9), S increased greatly to a value of 86%, indicating that the average decay in this condition was more nearly a straight line. In the check condition A10, where the diffusers had been replaced by equivalent absorbers, S had an intermediate value of 78%. The same pattern is evident in conditions B8, B9 and B10, where measurements were made in the room with non-parallel walls,

Table 5

Effect of Room Condition on Double-Slope Ratio,

Room condition	Frequency band				Mean	Stand error mea
	700 c/s	1 kc/s	1·4 kc/s	2 kc/s		
A8	62	69	68	70	67	1·
A9	91	83	85	86	86	1·
A10	75	80	71	87	78	1·
B8	75	74	67	67	71	1·
B9	75	82	86	84	82	1·
B10	79	79	76	80	79	1·
B8*	78	64	69	70	70	1·
B9*	77	80	80	88	81	1·
B10*	82	71	68	82	76	1·
A4	75	79	77	75	77	1·
A5	75	74	80	81	77	1·

* Absorbers and diffusers rearranged.

although the range of the means is appreciably less. When materials were rearranged in the same room and the meas ments repeated to provide conditions B8*, B9* and B10*, values of S agreed very well with those for conditions B8, and B10. The area of absorber in B10* was somewhat gre than that in B10.

The values for conditions A4 and A5 are also listed in Table. These values are not significantly different, indica that the curvature of the decays is not reduced when diffu are added to the empty room, where the absorption may regarded as evenly distributed. The curvature of the decay these two conditions was greater than in condition A9, where diffusers were present and the absorption in the room was b distributed. The apparent anomaly represented by this c parison remains unexplained. It must be noted, however, conditions A4 and A5 were part of a different set of experim from those relating to A9 and that the two sets differed wi in their average reverberation times. The reduction of reverb

on time between conditions A4 and A9 is sufficient to produce 1 appreciable alteration in the detailed appearance of the ecays, the average number of oscillations for a given level nange being reduced. It is thought that such differences in ppearance may affect the judgment of an observer in assigning n average slope to part of a decay and lead to systematic eviations where the effect is marked. A direct comparison etween A4 and A5 on the one hand and A9 on the other may nerefore not be justifiable.

(6) REVERBERATION-TIME CONTOURS

The magnitude of the parameter P in the foregoing Sections eferred to the spread of measured reverberation time in the oom but took no account of the manner in which the reverbera-on time varied throughout the room. The same magnitude ould be obtained when the readings of reverberation time hanged continuously across the room as in the case when there vere marked local variations.

In this experiment, reverberation time was measured in about 100 microphone positions at 1 ft intervals in one plane of a room with 10 ft sides. In the first instance, warble-tone pulses were radiated from a loudspeaker and contours of equal reverberation time were plotted on a plan of the room. With this technique the essential information contained in the diagram was masked by unnecessary detail, the form of which varied without apparent change in experimental conditions. The warble-tone technique was therefore discarded.

In the method which proved the most successful, the loud-speaker was supplied with pulses of white noise filtered through one-third-octave band-pass filters. In plotting the reverbera-tion-time contours the times chosen were running averages from groups of four adjacent microphone positions. It was found convenient to normalize contour levels in terms of percentages of the mean reverberation time because this simplified the com-parison of plots having different mean reverberation times.

Fig. 15 shows the contour plots for a horizontal plane at half

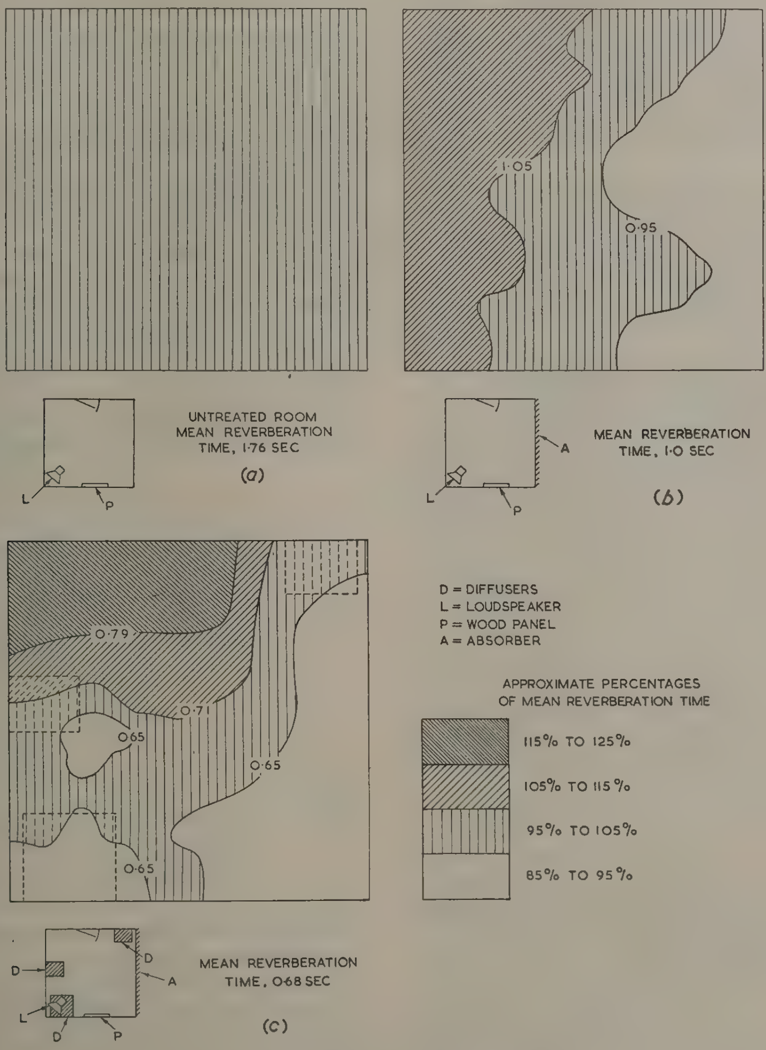


Fig. 15.—Reverberation-time contours measured in cubical room.
Experimental conditions appropriate to each plot are indicated in the plan diagrams of the room.

room height for the 1 kc/s filter setting. In the condition represented by Fig. 15(*a*) the room was untreated. No contours are shown, because the closest contours to the mean are at ±5% of it and there is no departure greater than 5% in this condition. This illustrates the substantial uniformity of values in the untreated room. In the condition represented by Fig. 15(*b*), one of the walls was covered with a medium-frequency absorber. The simple pattern indicates the systematic lowering of reverberation time towards the absorbing wall. In Fig. 15(*c*), which shows the results with three rectangular diffusers fixed to the untreated walls in or near the measurement plane, the pattern has been made more complex by the presence of the diffusers. In each of these diagrams the contour level is 10% of the mean reverberation time.

The work done up to the present is largely exploratory, and these diagrams are included only as examples of the results to be expected. The technique shows promise and may well be developed and applied to the study of broadcasting studios.

(7) EFFECT OF DIFFUSION ON EFFECTIVE ABSORPTION COEFFICIENT OF AN ABSORBER

Experience suggests that the efficiency of an absorber in a room is related to the degree of diffusion of the sound field in the room. Thus the measured absorption coefficient of a material might be used as an index of the degree of diffusion. Two experiments were carried out in which this hypothesis was examined.

The first experiment was designed to examine the effects of diffusers on the measured absorption coefficient of a material in the cubical room A. Twelve diffusers similar to those described in Section 2.4 were mounted on four of the surfaces of the room, while 100 ft^2 of a special medium-frequency absorber, adopted as a standard for these experiments, was mounted in one of the following two ways:

(*a*) Distributed on five surfaces (between the diffusers). The units were in several sizes, the largest area being 16 ft^2.

(*b*) Restricted to and completely filling one wall.

The absorption coefficient of the standard sample was determined by measuring reverberation times with and without the sample, and with and without the diffusers in the room; in this way the absorption directly due to the diffusers, in any case small, did not enter into the calculation of the absorption

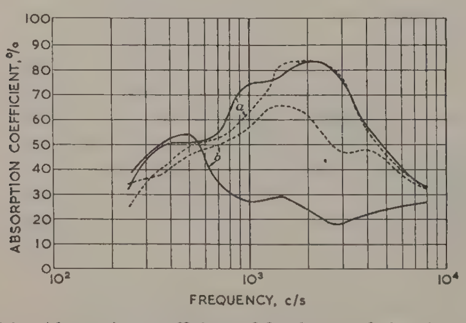


Fig. 16.—Absorption coefficient of fixed area of absorber under various conditions.

(*a*) Distributed throughout the room.
(*b*) Concentrated on one wall.
—— No diffusers.
---- With diffusers.

coefficient. The results are plotted in Fig. 16. Above 500 c/s a considerable difference in the absorption coefficient between arrangements (*a*) and (*b*) was observed when no diffusers were present, the absorption coefficient in the 2 kc/s region being reduced to about 25% of its former value when the absorb was concentrated on one wall. In the same condition, howev the absorption coefficient rose to about 75% of the original val when diffusers were present. In arrangement (*a*) where t absorber was distributed the diffusers had no effect.

It appears, therefore, that diffusers of this sort greatly impro conditions when the distribution of the absorber is extreme bad, but that there is no improvement when the absorber well distributed and therefore reaches its maximum efficien without additional diffusers.

The second experiment was intended to find whether, in vi of the equivalence of diffusers and absorbing patches found Section 2.6, distributing the absorption evenly was as effect as adding diffusers in increasing the efficiency of a concentrat absorber.

The conditions were similar to those of the first experime except that room B was used. The sample under considerati was restricted to one wall throughout the experiment, additio absorption being provided to produce a diffuse condition wh required. Two determinations of the absorption coefficient the sample were made by measuring reverberation time in fo conditions of the room:

(i) Empty.
(ii) With the standard sample of area 96 ft^2 concentrated one wall.
(iii) With the additional absorption distributed around room. No sample.
(iv) With additional absorption as in (iii) and sample co centrated on one wall as in (ii).

Two further conditions were used in order to compare effect due to distributed absorption with that due to rectangu diffusers in this room. These were:

(v) With 12 rectangular diffusers and no sample absorber.
(vi) With diffusers as in (v), with the sample restricted to o wall.

The absorption coefficients obtained in these conditions shown in Fig. 17. Curve (*a*) was obtained from conditions

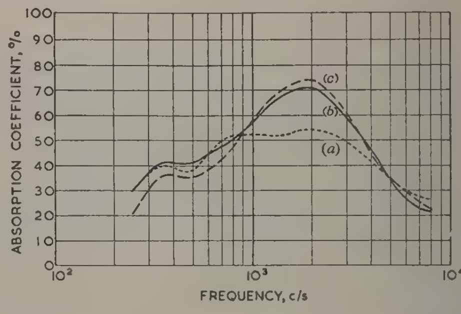


Fig. 17.—Absorption coefficient of absorber concentrated on on wall of room.

(*a*) Room otherwise empty.
(*b*) With additional absorber distributed around room.
(*c*) With rectangular diffusers, no additional absorber.

and (ii); curve (*b*) from conditions (iii) and (iv); and curve from conditions (v) and (vi).

It will be seen that in the region above 500 c/s the presence additional distributed absorption increased the measured abso tion coefficient by as much as 30%. This increase is, wit experimental error, the same as that obtained when rectangu diffusers were present instead of distributed absorption. T effect is smaller than that obtained for rectangular diffusers

the first experiment, presumably owing to the higher degree of diffusion associated with the non-parallel walls and a larger volume in the room used for the second experiment. In fact, the value obtained for the concentrated absorber in the empty room is considerably higher than the equivalent value in the first experiment, taking into account that the absorbers were not the same in the two cases. In both experiments the value obtained with rectangular diffusers was about 75% of that obtained with a distributed absorber (not plotted in Fig. 17).

(8) CONCLUSIONS

(8.1) Summary of Conclusions from Individual Experiments

In the foregoing work two interconnected problems have been studied. An attempt has been made to find a measure of the state of diffusion in a room, and the physical effects of varying the distribution of absorption and adding artificial diffusers have been investigated. To study these problems together it has been necessary to make certain general assumptions. The first is that the physical quantities studied are those likely to be related to the state of diffusion, and the second is that the room conditions are those that can reasonably be expected to increase or decrease the degree of diffusion.

The investigation became simplified to a series of experiments where room conditions were changed as markedly as possible, and the resulting alterations of each of the physical quantities were examined. Any quantity or parameter which responds to changes in the absorption and the shape of the room may offer a worth-while means of studying conditions in broadcasting studios and auditoria in general. The experiments show that it is not difficult to detect changes in the sound field as a result of altering the distribution of absorption or of adding artificial diffusers, but although evidence can be found for changes in several physical quantities, they are not necessarily of practical value for use in experimental investigations in the laboratory or in the field. In some cases a great deal of trouble had to be taken to be sure that a change had indeed taken place, even though the room changes were made as great as possible. The following general conclusions are reached.

Sound Decay Irregularity.—This parameter was mainly investigated with reference to the effect of diffusers. Great care is required to ensure accuracy and consistency in results; in particular the method is very sensitive to small instrumental variations. It is nevertheless to be noted that the magnitude of the sound decay irregularity was the only quantity which indicated an improvement in the degree of diffusion when diffusers were added to an empty room, although this indication was not pronounced.

Spatial Variation of Reverberation Time, P.—This quantity responds quite markedly to changes in the distribution of absorption. This is not surprising, even ignoring general considerations of diffusion, as one might expect to find below-average reverberation times in areas of a room close to surfaces where the absorption coefficient is high. It may well be useful in assessing the behaviour of broadcasting studios.

Frequency Variation of Reverberation Time, F.—There is no obvious reason to expect the variation of reverberation time with small increments of frequency, unlike the spatial variation, to depend on the position of absorbing surfaces. Even so, this quantity gave quite clear indications of changes in the distribution of absorption. It also gave a better indication of the presence of diffusers than the parameter P in a room with inherently poor diffusion. It is considered that this quantity may well be a very useful index for use in field measurements.

Variation of Reverberation Time with Orientation of Microphone.—Although some evidence of change of reverberation time with orientation of the microphone was observed with major changes in the distribution of absorption, little information could be gained under the conditions of these experiments. The evidence of the present work suggests that study of directional effects in small studios is not likely to be fruitful, although important work on similar phenomena in large rooms and concert halls has been reported by Meyer and Thiele.[5]

Sound Decay Curvature, S.—In the work described, the curvature was obtained in the simplest possible way by measuring the slopes of the first and second halves of the sound decay curve. It would be possible to extend the principle to derive more subtle measurements of curvature, but in the form here used, where one slope was expressed as a ratio of the other, it is considered to be a very useful index and one which has the advantage of simplicity. In the experiments it gave clear indications of the presence of bad distribution of absorption and of the effects of artificial diffusers.

Reverberation-Time Contours.—Contour diagrams of this kind do not provide a single numerical index of diffusion. The paper gives an account of the first exploratory steps in what may become a useful technique. Investigation is required to determine how to plot the contours so that essential information is retained and redundant matter discarded. If this can be done, the method will be useful for special investigation of broadcasting studios in which faults in the distribution of absorption are suspected.

(8.2) Discussion

In making these judgments of the effectiveness of certain techniques for investigating diffusion, it has been necessary to make assumptions regarding the qualitative effect of distribution of absorption and of artificial diffusers. The usefulness of the techniques described having been assessed, it is now possible to draw conclusions regarding the degree to which the state of diffusion is changed by variations in the room, and to compare the effects of distributed absorption and artificial diffusers.

The first experiments, using the indices P and F, showed that, when the absorption was restricted to one of the surfaces of a room, the poor degree of diffusion then apparent was markedly improved by a more uniform distribution of the absorbing surfaces over all surfaces in the room.

The presence of diffusers in the otherwise empty room did not appreciably change the degree of diffusion, as indicated by the spatial and frequency variation indices, P and F, or by the double slope index S. The earlier decay-irregularity experiments however had shown a significant improvement with the presence of diffusers.

The effect of diffusers was most marked when they were used for the treatment of a room where the absorption was badly distributed. All three of the indices mentioned above indicated a clear improvement in the degree of diffusion, making due allowance for absorption introduced by the diffusers themselves. The improvement was repeated when variations of the same basic experiment were carried out. It was observed that, where measurements were made in a somewhat larger room with non-parallel walls, the difference between treated and untreated conditions was less marked, indicating that the non-parallel walls had been effective in improving the degree of diffusion.

The experiments on the measurement of absorption coefficient, using this quantity as an indicator of diffusion conditions, again showed that, where a reasonable uniformity of distribution of absorption over the walls existed, diffusers had no effect on the measured value of absorption coefficient. However, where the absorption was concentrated on one wall, the presence of diffusers effected a marked increase in the measured absorption

coefficient. A similar effect on the measured absorption coefficient of this concentrated sample was observed when additional absorption was arranged on the other surfaces of the room. When linked with other experiments, not described here, it was clear that the maximum absorption coefficient measured under these conditions was effectively the maximum absorption coefficient attainable for the material.

Results confirming these conclusions on the relationship between diffusion and the efficiency of absorbers have been recently published by Meyer and Kuttruff.[6] From results with models, they show that the effective absorption coefficient of a material fixed to the walls of an enclosure may be regarded as a measure of the state of diffusion, and using this criterion they discuss the effect of hemicylindrical diffusers and room shape.

(8.3) General Conclusions

Several broad conclusions arise from the work described in the paper. The principal conclusion is that practical means are available for the measurement of some, at least, of the characteristics of good sound diffusion. These are likely to be of value for the investigation of the acoustics of sound studios and auditoria and in the measurement of absorption coefficients.

The method which appears to be most suitable for adoption as a routine criterion of diffusion in operational studios is that based on the measurement of the average curvature of the individual decay curves. This method is simple and gives a satisfactory indication of changes in diffusion from all causes.

The methods of investigation taken together show that the degree of diffusion may be improved by the presence of rectangular diffusers and by suitable arrangement of absorbing material, and there is some experimental evidence to confirm the hypothesis that diffusers will effect an improvement even in a room with surfaces of uniform absorption coefficient.

(9) ACKNOWLEDGMENTS

The authors wish to thank the Chief Engineer of the British Broadcasting Corporation for permission to publish the paper. They also wish to acknowledge the considerable assistance of their colleagues in the B.B.C. Research Department.

The results and conclusions from the experiment describ in Section 7 were presented by Mr. T. Somerville of the B.B Research Department, at the International Organization Standardization Conference in Paris, January, 1957.

(10) REFERENCES

(1) SOMERVILLE, T., and WARD, F. L.: 'Investigation of Sou Diffusion in Rooms by means of a Model', *Acust* 1951, **1**, p. 40.
(2) EYRING, C. L.: 'Reverberation Time in Dead Room *Journal of the Acoustical Society of America*, 1930, p. 217.
(3) BOLT, R. H., and ROOP, R. W.: 'Frequency Respo Fluctuations in Rooms', *ibid.*, 1950, **22**, p. 280.
(4) SOMERVILLE, T.: 'An Empirical Acoustic Criteri *Acustica*, 1953, **3**, p. 365.
(5) MEYER, E., and THIELE, R.: 'Raumakustische Untersu ungen in Zahlreichen Konzertsälen und Rundfunkstud unter Anwendug Neuerer Messverfahren', *ibid.*, 1956 (Beiheft 2), p. 425.
(6) MEYER, E., and KUTTRUFF, H.: 'Akustische Modellversu zum Aufban eines Hallraumes', *Nachrichten der Akade der Wissenschaften in Göttingen*, 1958, **6**, p. 97.
(7) MORSE, P., and BOLT, R. H.: 'Sound Waves in Room *Reviews of Modern Physics*, 1944, **16**, p. 137.
(8) RETTINGER, M.: 'Sound Diffusion in Rooms', *Au Engineering*, August, 1954, **38**, p. 19.
(9) BOLT, R. H.: 'Normal Frequencies in Rectangular Room *Journal of the Acoustical Society of America*, 1952, p. 80.
(10) SOMERVILLE, T., and GILFORD, C. L. S.: 'Acoustics of La Orchestral Studios and Concert Halls', *Proceedings I.E.* Paper No. 2297 R, March, 1957 (**104** B, p. 85).
(11) BRUEL, P. V.: 'Automatic Recording of Frequency Irre larity in Rooms', *Acustica*, 1954, **4**, p. 21.
(12) LAMORAL, R.: 'Sur le coéfficient d'absorption des matéri acoustiques', *Onde Électrique*, 1953, **33**, p. 461.
(13) LAMORAL, R., and TREMBASKY, R.: 'Nouvelle car térisque acoustique des salles et réalisation d'un compt électronique polyvalent', *ibid.*, 1956, **36**, p. 441.

621.375.422

The Institution of Electrical Engineers
Paper No. 3227 M
Mar. 1960

A LOW-DRIFT TRANSISTOR CHOPPER-TYPE D.C. AMPLIFIER WITH HIGH GAIN AND LARGE DYNAMIC RANGE

By I. C. HUTCHEON, M.A., A.M.I.Mech.E., Associate Member, and D. SUMMERS.

(*The paper was first received* 26*th June*, 1959, *and in revised form* 9*th January*, 1960. *It was published in March*, 1960, *and was read before the* MEASUREMENT AND CONTROL SECTION 15*th March*, 1960.)

SUMMARY

The paper describes a transistor chopper-type d.c. amplifier which is intended for use in the field of process measurement and control. It has a voltage drift below $\pm 10\,\mu$V and a current drift below $\pm 4 \times 10^{-9}$ amp.

The forward gain in the steady state is about 10 kV/μA and 15 volts/μV, and, in consequence, the full output swing of 0 to 5 volts is provided by an input signal little greater than the drift. The application of sufficient overall negative d.c. feedback therefore enables input signals of 0–10 mV or 0–4 μA to provide full output with an accuracy of $\pm 0 \cdot 1$%.

The low drift is obtained by operating the transistor chopper at 200 c/s, stabilizing its temperature within ± 2°C, and providing suitably stable waveforms to drive the transistor and compensate for its voltage offset.

The a.c. gain prior to the demodulator is only 50 volts/μA, the remaining voltage gain, of about 200, being provided by a low-drift d.c. amplifier which is connected as a feedback integrator and used to smooth the demodulated output. This arrangement enables the system to handle, with a minimum of saturation, the large error signals which occur during a rebalancing operation, and thus maintain a fast response to large changes of input signal.

The system is first analysed in general terms, and expressions are derived which describe its performance and act as a guide to the design of this type of amplifier.

LIST OF SYMBOLS

A = Voltage gain of d.c. amplifier.
A_i = Steady-state voltage gain of integrator.
C = Integrating capacitance, farads.
C_1 = Input coupling capacitance, farads.
C_2 = Output coupling capacitance, farads.
C_J = Input filter capacitance, farads.
F = Maximum input current $\times\, nZ/V_t$.
I = Direct input current, amp.
I_f = Direct feedback current, amp.
I_i = Current drift due to integrator, referred to input, amp.
I_J = Current drift of input chopper (transient effects excepted), amp.
I_n = Current offset due to Q_n, amp.
I_p = Current offset due to Q_p, amp.
i = Current drift of d.c. amplifier, amp.
n = Ratio of unclamped to clamped period.
Q_n = Charge of input-chopper negative transient, coulombs.
Q_p = Charge of input-chopper positive transient, coulombs.
R = Integrator smoothing resistance, ohms.
R_f = Feedback resistance, ohms.
R_{in} = Input impedance of a.c. amplifier, ohms.
R_J = Input-filter terminating resistance, ohms.
R_L = Integrator load resistance, ohms.
R_o = Output impedance of a.c. amplifier, ohms.
R_s = Source resistance, ohms.
r_o = Resistance of input chopper when open, ohms.
r_s = Resistance of input chopper when closed, ohms.
T = Time-constant in linear régime, sec.
τ = Duration of clamped period, sec.
V_i = Voltage drift due to integrator, referred to input chopper, volts.
V_J = Voltage drift of input chopper (transient effects excepted), volts.
V_{max} = Limiting amplitude of a.c. amplifier output, volts.
V_o = Integrator output voltage, volts.
V_{omax} = Maximum value of V_o, volts.
V_p = Voltage offset due to Q_p, volts.
V_r = Amplitude of output ripple, volts.
V_{rs} = Value of V_r in the steady state, volts.
V_s = Voltage drift of integrator, volts.
V_t = $V_s + \frac{1}{2}V_{omax}/A_i$, volts.
v = Voltage drift of d.c. amplifier, volts.
Z = Transfer gain of a.c. amplifier, volts/amp.

(1) INTRODUCTION

Since the forward gain of a chopper-type d.c. amplifier cannot usually be held within close limits, sufficient negative d.c. feedback must be applied over the whole amplifier to define a much lower net gain with whatever accuracy is required.

Consider, for example, the half-wave arrangement shown in Fig. 1, in which the resistor R_f provides the feedback. If the

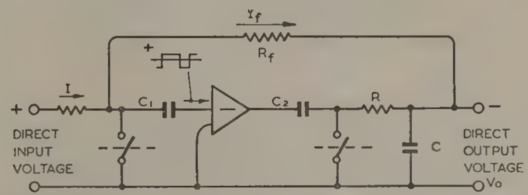


Fig. 1.—Half-wave chopper-type d.c. amplifier with *RC* smoothing.

net gain is required to remain constant within $0 \cdot 1$%, the steady-state error signal applied to the amplifier must not exceed $0 \cdot 1$% of the input current I, and the forward gain must be at least 1 000 times the gain of the feedback path, i.e. $1\,000 R_f$ volts/amp. The output voltage V_o is then very nearly equal to IR_f, and, if a small input current is to provide a large output, both R_f and the forward gain must be large.

As R_f and the forward gain are increased, the error signal decreases until eventually the drift of the amplifier becomes the significant factor and prevents further reduction of the error. Any further increase in gain gives no advantage, and the greatest gain which can usefully be employed is that which makes the total error signal little greater than the drift.

A difficulty arises, however, if all the gain is embodied prior to the demodulator. Any substantial change in the input signal causes the amplifier to receive an error signal which is temporarily much greater than the steady-state value. The amplifier

Mr. Hutcheon and Mr. Summers are with George Kent, Ltd.

and demodulator saturate and deliver a severely limited output, the smoothing capacitor charges at a very restricted rate and the system has a poor response to large changes of input signal. Furthermore the system is easily rendered insensitive by the saturating effect of spurious input signals, such as, for example, 50 c/s pick-up in long input cables from a distant source.

Both disadvantages are overcome if a sufficient part of the total gain is provided by a d.c. amplifier after the demodulator, as shown in Fig. 2(*a*). The a.c. gain prior to the demodulator

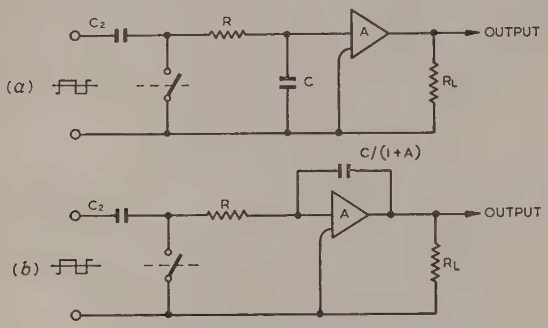


Fig. 2.—Alternative output stages.
(*a*) *RC* integrator plus d.c. amplifier.
(*b*) Feedback integrator.

is then much lower, demodulation is performed at a relatively low signal level and saturation is avoided.

The d.c. gain which can usefully be employed depends on the drift of the d.c. amplifier, and it is not always possible to prevent saturation effects altogether. Nevertheless the arrangement permits much more of the operation to take place in the linear régime, and the response to large changes of input is much better than if no d.c. amplifier were provided.

Two further advantages are gained if the *RC* integrator and d.c. amplifier are converted into an equivalent feedback integrator, as shown in Fig. 2(*b*). First, the capacitance, which otherwise may be very large in high-gain systems, is reduced by a factor $1 + A$ to a more convenient value. Secondly, the output is almost unaffected by ripple or random fluctuations in supply voltages, since it is held constant by feedback action through the capacitor.

Operation of the half-wave chopper-type d.c. amplifier with a feedback integrator smoothing circuit is analysed in the following Section, and the results are then applied to the design of a practical amplifier.

(2) ANALYSIS OF THE HALF-WAVE CHOPPER-TYPE D.C. AMPLIFIER WITH FEEDBACK INTEGRATOR

Fig. 3(*a*) is a block diagram of the complete amplifier which comprises an input chopper S_1, a capacitance-coupled amplifier having a transfer gain of Z volts/amp, an output chopper (demodulator) S_2 and a feedback integrator. The switches are assumed to open and close in synchronism, both being open for a time $n\tau$ and closed for a time τ during each cycle. Negative d.c. feedback is applied over the whole system, and it is assumed in the analysis that this defines the output voltage in terms of the input current. Any of the four possible feedback combinations can equally well be used, however, with little change in the argument. Typical waveforms are shown in Figs. 3(*b*)–(*e*) and refer to operation with the feedback disconnected.

(2.1) Operation

Once in every cycle, the two switches close simultaneously for a time τ, and the coupling capacitors C_1 and C_2 have their

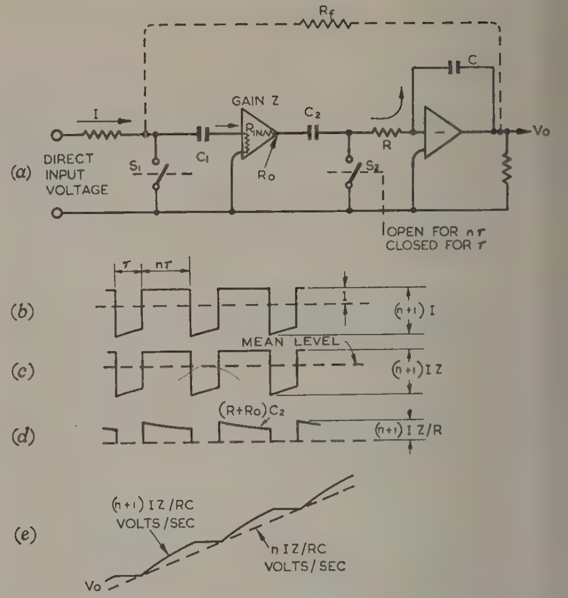


Fig. 3.—Half-wave chopper-type d.c. amplifier with feedback integrator.

(*a*) Circuit.
(*b*) A.C. amplifier input current.
(*c*) A.C. amplifier output voltage.
(*d*) Integrator input current.
(*e*) Integrator output voltage.

potentials reset to values which ensure that zero input sign corresponds to zero output. During the periods $n\tau$ when th switches are open, therefore, the amplifier and its two capacito form a drift-corrected d.c. amplifier connected between the inp and the integrator, and if any error signal is present, the outpu changes in a direction which tends to reduce it. The tim constant of the input circuit in the clamped state is made ju sufficient to preserve a moderately square waveform, while th of the output circuit is made shorter than the clamped perio so that the resetting process is carried as far as possible durin each cycle. Both time-constants in the unclamped state a made relatively long.

The system has three different modes or régimes of operatio which have to be considered separately. These are:

(*a*) The saturated régime, in which a large error signal causes th demodulator to deliver its maximum output and the output of th integrator changes at a steady rate.

(*b*) The linear régime, which obtains when the error signal reduced to a value which no longer saturates the demodulator an in which the response is exponential.

(*c*) The steady state, which exists when a state of balance h been reached and the error signal is just sufficient to counteract th drifts of the input chopper and the integrator and to provide th necessary change in output from its mean level.

The performance in each of these régimes and the demarcatio between them are deduced from the open-loop characteristics o the amplifier, which are therefore determined first.

(2.2) Open-Loop Characteristics

Consider the system shown in Fig. 3(*a*) with the overall d. feedback disconnected and a direct current I, which does no saturate the amplifier or the demodulator, applied suddenly t the input terminals.

During the first few cycles, the potential across C_1 rises to value nIR_{in}, which causes the current I flowing into the ampl

er during the unclamped periods, $n\tau$, to be balanced by a urrent nI flowing out during each clamped period τ. The input hopper thus has a current gain which rises from unity to $n+1$ fter a short delay. The potential across the chopper during e unclamped periods rises from IR_{in} to $(n+1)IR_{in}$ during the me few cycles, and the input impedance is increased to $(n+1)R_{in}$.

Thus, after a few cycles, the amplifier receives a moderately ectangular current waveform of amplitude $(n+1)I$ as shown Fig. 3(*b*), and delivers a corresponding output voltage of mplitude $(n+1)IZ$ as shown in Fig. 3(*c*).

The output chopper holds the right-hand side of C_2 at earth otential during each clamped period. Since R_oC_2 is low, the .c. restoration is substantially complete and direct-voltage ulses of amplitude $(n+1)IZ$ are applied to R during each nclamped period, causing current pulses $(n+1)IZ/R$ to flow nto C, as shown in Fig. 3(*d*). The mean voltage applied to over a complete cycle is nIZ, and the input chopper, a.c. mplifier and output chopper taken together can be regarded as aving a mean transfer gain

$$nZ \text{ volts/amp} \quad \ldots \ldots \quad (1)$$

hich is achieved after a delay of a few cycles. The integrator utput voltage therefore changes at a mean rate

$$\frac{dV_o}{dt} = \frac{nIZ}{RC} \quad \ldots \ldots \quad (2)$$

hich, as shown in Fig. 3(*e*), is composed of ramp increases t a rate $(n+1)IZ/RC$ volts per second during the unclamped eriods alternated with pauses during the remainder of each ycle. Inspection of Fig. 3(*e*) shows the amplitude of the ripple ontent of the output voltage to be approximately

$$V_r \simeq n\tau IZ/RC \quad \ldots \ldots \quad (3)$$

(2.3) Saturated Régime

If the feedback is connected and the input current undergoes a large step change, the a.c. amplifier delivers an output voltage whose amplitude is limited by the circuit parameters to some value V_{max}. (In principle either the amplifier or the demodulator may be the limiting factor, but the effect is the same.) The output voltage V_o of the integrator then changes at a rate V_{max}/RC during each unclamped period, remaining stationary as before while the output chopper is clamped, and its mean rate of change is

$$\left(\frac{dV_0}{dt}\right)_{max} = \frac{n}{n+1}\,\frac{V_{max}}{RC} \quad \ldots \ldots \quad (4)$$

The output continues to change at this rate until the feedback has reduced the net input current to a value which just provides the limiting a.c. output. The a.c. output then falls and the remainder of the rebalancing action takes place in the linear régime. The transition occurs when

$$(I - I_f)(n+1)Z = V_{max}$$

i.e. when

$$I - I_f = \frac{V_{max}}{(n+1)Z} \quad \ldots \ldots \quad (5)$$

The maximum current which the a.c. amplifier has to supply during the unclamped periods is V_{max}/R in either direction. If $n > 1$, however, the amplifier must be capable of supplying a greater current nV_{max}/R (also in either direction) while the output chopper is clamped, in order to reset the potential of C_2 completely.

(2.4) Linear Régime

The last part and, in some cases the whole, of a rebalancing action takes place with the a.c. amplifier unsaturated. If the action covers more than a very few cycles, it is reasonable to assume that the input chopper has a gain of $n+1$, in which case the system reduces to that shown in Fig. 4(*a*).

Suppose, therefore, that a step change of input current I is applied to the system shown in Fig. 4(*a*) at time $t = 0$. After

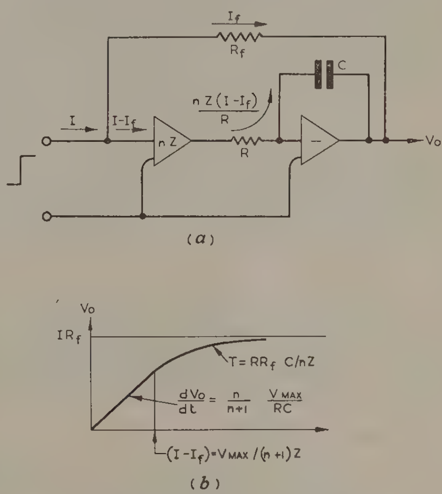


Fig. 4.—Response with overall d.c. feedback.
(*a*) Equivalent circuit.
(*b*) Output, amplifier saturated initially.

t seconds the feedback current will have risen to a value I_f, the net input is $I - I_f$, and the current flowing into the integrating capacitor is $nZ(I - I_f)/R$. Therefore

$$\frac{dV_0}{dt} = \frac{nZ(I - I_f)}{RC} = R_f\frac{dI_f}{dt} \quad \ldots \ldots \quad (6)$$

and
$$I_f = I(1 - \varepsilon^{-t/T}) \quad \ldots \ldots \quad (7)$$

where
$$T = \frac{RR_fC}{nZ} \quad \ldots \ldots \quad (7a)$$

and
$$I_f = 0 \text{ when } t = 0 \quad \ldots \ldots \quad (7b)$$

Thus I_f approaches I and V_o approaches IR_f in an exponential manner. This is shown in Fig. 4(*b*), which includes also a portion of the response to a saturating signal.

If the time-constant calculated from eqn. (7*a*) is comparable with the cycle duration, the gain of the input chopper cannot be assumed equal to $n+1$ and the equation is invalid. The gain is never less than unity, however, and the time-constant during each unclamped period is therefore never greater than

$$T' = \frac{RR_fC}{Z} \quad \ldots \ldots \quad (7c)$$

(2.5) Steady State

When a state of balance has been reached, there remains a small steady-state error signal which, when amplified, is just sufficient to hold the integrator output voltage at the required steady value and counter any drifts arising in the input chopper. The feedback signal is assumed to be sufficiently free from ripple that the error signal can be regarded as constant, and the error comprises several components which are calculated as follows.

(2.5.1) Errors due to Integrator.

The integrator is initially adjusted so that its output voltage is half the maximum value when the applied voltage is zero. Changes in ambient temperature and ageing effects cause the d.c. amplifier to drift, however, and in order to maintain the same output, a voltage

$$V_s = v + iR \quad . \quad . \quad . \quad . \quad . \quad . \quad (8)$$

must be applied to the integrator, where v and i are the voltage and current drifts of the d.c. amplifier referred to its input. The corresponding current drift at the input to the system is equal to V_s divided by the mean transfer gain preceding the integrator. Thus

$$I_i = \frac{V_s}{nZ} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (9)$$

In addition to I_i, a small input current must be provided to swing the integrator output by a maximum of $\pm\frac{1}{2}V_{omax}$ about its mean value. If the steady-state gain of the integrator is A_i, the current has a maximum value

$$\pm\frac{\frac{1}{2}V_{omax}}{nZA_i} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (10)$$

Multiplying the above currents by R_s plus the input impedance $(n+1)R_{in}$ gives the associated voltage changes at a source of resistance R_s.

(2.5.2) Total Errors.

The input chopper contributes a current drift I_J and a voltage drift V_J. If the system is arranged for current amplification, therefore, the maximum error between the input and feedback currents is

$$I_i \pm \frac{\frac{1}{2}V_{omax}}{nZA_i} + I_J \quad . \quad . \quad . \quad . \quad . \quad (11)$$

In the case of voltage amplification, both V_J and the voltage drift caused by I_J in flowing through the source resistance must be added to the voltage drift due to the integrator, and the total voltage error is

$$\left(I_i \pm \frac{\frac{1}{2}V_{omax}}{nZA_i}\right)[R_s + (n+1)R_{in}] + I_J R_s + V_J \qquad (12)$$

(2.5.3) Output Ripple.

The mean voltage applied to the integrator in the steady state is within the limits $V_s \pm \frac{1}{2}V_{omax}/A_i$, and its greatest value is

$$V_t = V_s + \tfrac{1}{2}V_{omax}/A_i \quad . \quad . \quad . \quad . \quad . \quad . \quad (13)$$

It is applied as a series of direct-voltage pulses whose duration is $n\tau$ and whose amplitude is therefore $(1 + 1/n)V_t$, as shown in Fig. 5(*a*). No signal is applied during the clamped periods, and the output ripple is equal to the drift

$$V_{rs} = \frac{\tau V_t}{RC} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (14)$$

which then occurs. This is shown in Fig. 5(*c*). Since the drift is away from the correct value, it is necessary to make it smaller than the error signal, as was assumed initially.

(2.6) Deductions

(2.6.1) Gain of A.C. Amplifier.

The gain must be sufficiently great to reduce the integrator drift, referred to the input of the system, to a suitably small value. Thus, from eqn. (9),

$$Z \geqslant V_s/nI_i \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (15)$$

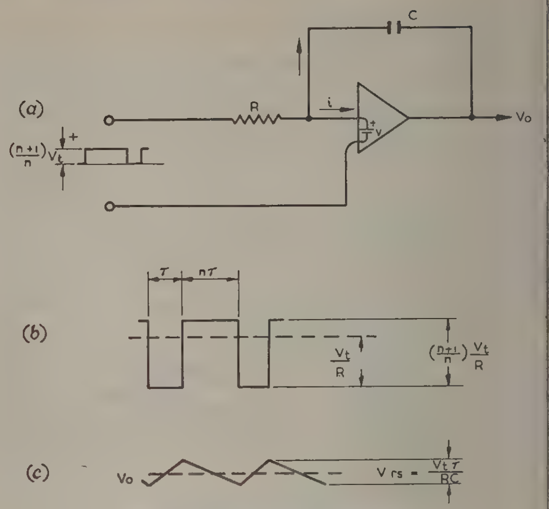


Fig. 5.—Steady-state conditions.
(*a*) Integrator with drift-correcting signal.
(*b*) Current in integrator capacitor.
(*c*) Output ripple V_{rs}.

where I_i is specified. A larger value of Z will improve the response in the linear régime, but is generally to be avoi since it increases the possibility of saturation by, for exam 50 c/s pick-up, and it does not improve the response in saturated régime. Thus Z should be held stable by a.c. feedb within the a.c. amplifier.

(2.6.2) Ratio of V_{max} to V_t.

The system is designed so that the input current V_t/nZ wh is required to drive the integrator under the worst condition a small fraction $1/F$ (e.g. 1/2 000) of the maximum input curr The application of maximum input would therefore provid mean demodulated output of FV_t if saturation did not oc The corresponding output from the a.c. amplifier is $(1 + 1/n)$ and there is no saturation if this is less than V_{max}, i.e. if

$$\frac{V_{max}}{V_t} \geqslant \left(1 + \frac{1}{n}\right)F \quad . \quad . \quad . \quad . \quad .$$

This cannot always be achieved in practice, but V_{max} is m large and V_t small, so that as much of the operation as poss takes place in the linear régime.

(2.6.3) Value of V_{max}.

V_{max} is determined in practice by the voltage rating of transistor used as the output chopper. Under saturated c ditions the collector receives a potential of either $+V_{max}$ $-V_{max}$ during the unclamped periods, according to the p of the a.c. signal. It must therefore be cut off by a positive voltage at least equal to $+V_{max}$, and consequently it m withstand at least $2V_{max}$ between base and collector when latter is negative. The transistor is desirably a symmetr type having a reasonable gain in both directions, and the grea base–collector voltage rating presently available is 30 volts. limits V_{max} to 15 volts and a suitable design value is 10 volt

(2.6.4) Value of R.

It is economic to make $R \simeq v/i$, thus equalizing the two c ponents of the integrator drift voltage V_s.

Reducing R below this value gives diminishing returns reducing V_s, while at the same time increasing the cur nV_{max}/R, which the a.c. amplifier must be able to supply du

ıe clamped periods. Increasing R above v/i, on the other hand, undesirable since it increases V_s.

Leakage through the integrating capacitor contributes to i ıd should be taken into account if substantial. Voltage errors ı the output chopper contribute to v but are generally negligible ı practice. Collector leakage of the output chopper in the ıt-off state has little effect since the output impedance of the .c. amplifier is low.

.6.5) Value of C.

The feedback capacitance, C, is chosen so as to make the eady-state output ripple substantially smaller than the perıissible error in the output signal, e.g. less than 1% or 0·1% of ıe maximum output. Thus, from eqn. (14),

$$C = \frac{\tau V_t}{R V_{rs}} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (17)$$

/here V_{rs} is the permissible output ripple. In systems handling ery small signals, noise may predominate over ripple and its eduction may demand a larger capacitance.

2.6.6) Values of n and τ.

When the input chopper is a transistor, transients due to the witching waveform give rise to a current offset which increases s $n\tau$ is reduced. This effect is described in Section 3.5 and nay determine a minimum value for $n\tau$.

In designs having no d.c. amplification after the demodulator, is sometimes made as great as 5 or 10, in order that τ and ıence the steady-state output ripple may be small. This enables to be reduced and improves the response proportionately.

Any increase in n, however, requires a proportionate increase n the current nV_{max}/R which the a.c. amplifier must be able to upply, and the use of large values of n appears to be a somevhat uneconomic method of improving the response. The use of a feedback integrator, on the other hand, may improve the esponse by as much as 100 or even 1 000 times, in which case a ınity mark/space ratio, which allows somewhat simpler circuits, s probably to be preferred.

In cases where C is determined by considerations of noise, educing τ does not allow C to be reduced, and a value of n arger than unity gives only the slight advantage that the ınclamped time is increased, thus allowing slightly increased lrive to the integrator.

2.6.7) Response with Perfect Integrator.

It is interesting to note the effect of using a perfect feedback ntegrator based on a d.c. amplifier having infinite gain and zero lrift. The output voltage would not change at all during the clamped periods, and C could be made infinitesimally small. Response to an error signal would therefore be infinitely fast, and the error would be reduced to zero immediately both switches were opened. The output signal would therefore follow the input signal exactly during the unclamped periods, but there would be a time delay of maximum value τ before input changes occurring during the clamped periods could be followed. This is the best response which a half-wave system can provide under any circumstances.

Although this ideal can never be reached, and a chopper system would be unnecessary if it could, the use of a feedback integrator does permit a good response to be obtained with high-gain systems using low chopping frequencies. It therefore seems particularly applicable when the frequency-dependent current drift of transistor choppers presents a problem, or where it is necessary to use a mechanical switch or other inherently low-frequency modulator in order to obtain a lower drift than transistor choppers can provide.

(3) DESIGN DETAILS

(3.1) Specification

One of the more important applications of d.c. amplifiers in industry is the amplification of thermocouple voltages. An accuracy of about $\pm 1°$C is commonly required, and this corresponds to a signal of about $\pm 50\,\mu$V if base-metal thermocouples are used, or ± 10 to $\pm 20\,\mu$V in the case of the platinum/platinum-alloy thermocouples used at the higher temperatures.

An amplifier for this type of application should therefore be capable of providing full output for an applied error signal not exceeding about $10\,\mu$V. Long runs of compensating cable together with a filter to reduce pick-up effects may create a source resistance of a few hundred ohms, and the current drift should therefore not exceed about 10^{-8} amp.

The accuracy should be maintained in ambient temperatures up to at least 50° C, and the amplifier, should continue to operate with reduced accuracy up to 60 or 70° C for short periods. The system should be unaffected by substantial series or common-mode 50 c/s pick-up.

(3.2) Stabilization of Input Chopper

In order to obtain the desired freedom from drift, the input chopper circuit, described by Chaplin[1, 2] and shown in Fig. 6(a),

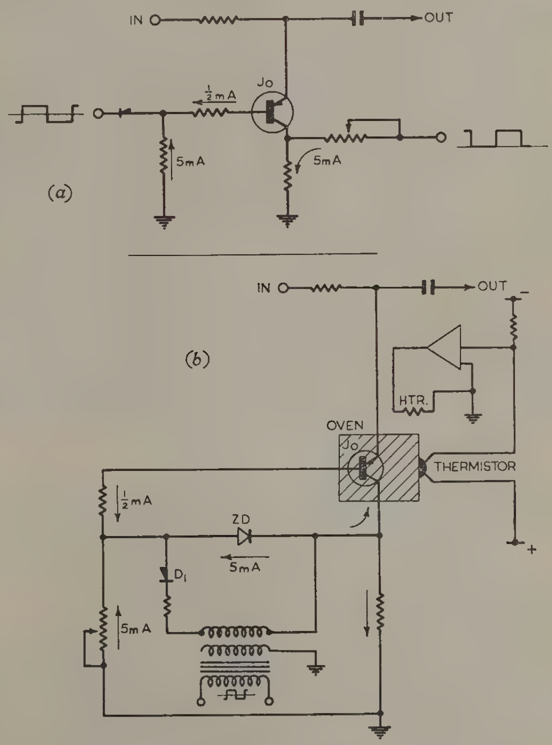


Fig. 6.—Input chopper.
(a) Basic arrangement.
(b) Circuit with bridge compensation and temperature control.

is modified according to Fig. 6(b), and a small oven is added which stabilizes the temperature of the transistor to within less than $\pm 2°$C at a nominal temperature of 50° C. The oven is a cylindrical aluminium block drilled to take the transistor, and it has a thermistor attached to one end and a heating coil wound on the outside. The thermistor and a fixed resistor are connected in series across a centre-tapped d.c. supply, and the current

derived from their junction controls the heater via a simple transistor d.c. amplifier.

The two circuits operate in substantially the same manner. A stabilized base current drives the transistor on in each case, a stabilized voltage compensates for the voltage offset, and a silicon diode, D_1, prevents any positive base drive when the transistor is off. The transistor is used with the collector earthed, since this arrangement reduces the voltage offset with typical unsymmetrical transistors. The modified arrangement has the advantage that a single Zener diode, ZD, stabilizes both the base current and the compensating voltage, whereas the other circuit requires two stabilized a.c. supplies of opposite phase. The use of a single winding with neither end earthed can give rise to additional transient effects due to inter-winding capacitances, but these are avoided by aligning the output winding with the other windings in such a manner that one end remains naturally at earth potential.

(3.3) Voltage Drift of Input Chopper

The voltage drift of the input chopper shown in Fig. 6(*b*) has the following main components:

(*a*) If the base current is chosen to suit the average transistor of a particular type, but not adjusted for each individual, the collector–emitter voltage changes by less than about $\pm 3\,\mu$V/deg C, and the drift is less than $\pm 6\,\mu$V when the temperature is held within $\pm 2°$ C.

(*b*) If the temperature coefficient of breakdown voltage of the Zener diode is below $\pm 0{\cdot}004\%$ per deg C, and the transistor offset is 2 mV, a change of $\pm 25°$ C in ambient temperature causes the compensating voltage to alter by less than $\pm 2\,\mu$V. This figure could be reduced to $\pm\frac{1}{2}\,\mu$V by using a more expensive Zener diode with a temperature coefficient below $\pm 0{\cdot}001\%$ per deg C. The effect of finite slope resistance in the Zener diode is rendered negligible by stabilizing the square-wave supply within a few per cent.

(*c*) The voltage between the base and emitter of the transistor has a temperature coefficient of about $-2{\cdot}5$ mV per deg C (germanium transistors) or $-1{\cdot}75$ mV per deg C (silicon), and temperature variations of $\pm 2°$ C give rise to changes of ± 5 mV or less. Since the Zener diode supplies about 5 volts with a stability of $\pm 0{\cdot}1\%$, the transistor base current remains stable within about $\pm 0{\cdot}2\%$, and the resultant drift in the collector–emitter voltage is negligible.

The total voltage drift, V_J, of the input chopper can thus be held below about $\pm 8\,\mu$V in the worst case, the effects of current drift in causing additional voltage drift being relatively negligible.

(3.4) Current Drift of Input Chopper

There are two sources of current drift:

(*a*) When the transistor is in the off condition, about one-tenth of the reverse leakage current of the diode D_1 reaches the base, while (in unsymmetrical transistors) less than one-tenth of this amount emerges from the emitter and constitutes a drift. Silicon diodes are now available which have a reverse leakage below about $10^{-3}\,\mu$A at 60° C, and enable the resultant drift to be held below 10^{-11} amp. Similar performance can also be obtained by using two silicon diodes of lower quality, one relieving the other of nearly all its reverse voltage.

(*b*) During successive clamped periods the input coupling capacitor charges to the residual drift voltage V_J of the transistor switch, as shown in Fig. 7(*a*). The voltage drift is absent during the unclamped periods, and if the integrator is to receive no drive, a direct current V_J/r_o must be applied in order to develop a voltage V_J across the open resistance r_o of the transistor. The

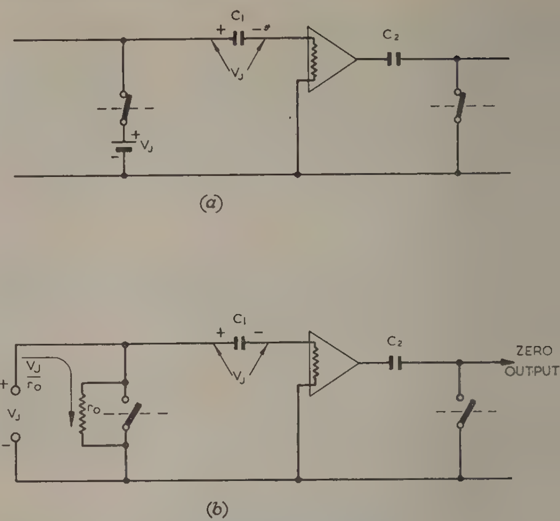


Fig. 7.—Effect of voltage drift.
(*a*) Clamped period τ.
(*b*) Unclamped period $n\tau$.

effect is shown in Fig. 7(*b*), and V_J/r_0 is a current drift. I quite negligible if a silicon transistor is used, since the value r_o at 50° C is then in the megohm region, but it may be su stantial if the transistor is germanium. The minimum value r_o for the type SB101 transistor at 50° C, for example, is abo 4 kilohms, and a voltage drift of $\pm 8\,\mu$V gives rise to a curr drift not exceeding $\pm 2 \times 10^{-3}\,\mu$A. The total current drift the input chopper, I_J, can thus be held below $\pm 2 \times 10^{-3}\,\mu$A a suitable germanium transistor is used. The operating te perature could be increased and the drift reduced by the use o silicon transistor, but with some increase in cost.

(3.5) Effects of Input Chopper Transients

The square-wave voltage existing at the base of the chop transistor, together with the base–emitter capacitance, ca

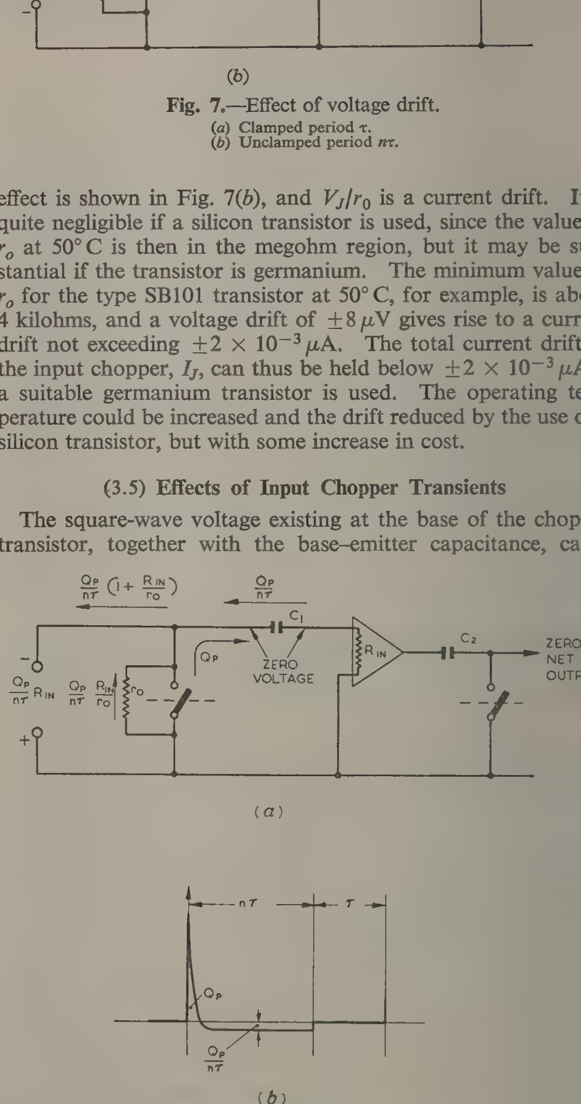


Fig. 8.—Effect of positive transient.
(*a*) Unclamped period $n\tau$.
(*b*) Current in C_1.

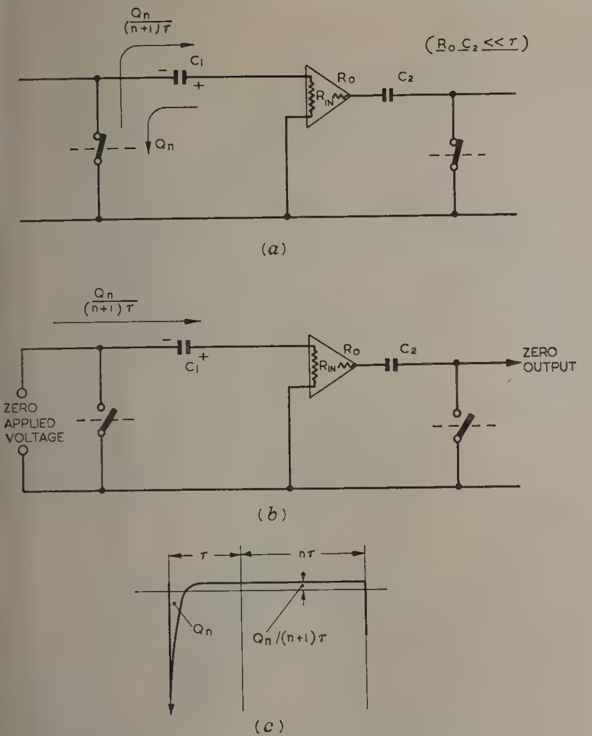


Fig. 9.—Effect of negative transient.

(*a*) Clamped period τ.
(*b*) Unclamped period $n\tau$.
(*c*) Current in C_1.

alternate positive and negative current transients to flow out of the emitter every time the transistor is switched off and on, the positive transient being increased by hole-storage effects in the base. Both transients, whose charges are denoted Q_p and Q_n, give rise to errors which are additional to the drift effects already described, and which can be assessed independently of one another. Their effects are shown in Figs. 8 and 9.

If the output chopper operates in exact synchronism with the input chopper, the amplified positive transients flow into the integrator and provide a false output. Assuming that they are not clipped by the a.c. amplifier, their effect can be exactly counterbalanced by a negative direct current $Q_p/n\tau$ drawn from the input of the a.c. amplifier during the whole of each unclamped period, as shown in Fig. 8(*a*). C_1 and C_2 then remain uncharged, and the corresponding voltage which must be applied across the input chopper is

$$V_p = -\left(\frac{Q_p}{n\tau}\right) R_{in} \quad . \quad . \quad . \quad . \quad . \quad (18)$$

This is the voltage offset due to Q_p. Since a current V_p/r_o is drawn through the open resistance r_o of the chopper, the total current offset due to Q_p is

$$I_p = -\frac{Q_p}{n\tau}\left(1 + \frac{R_{in}}{r_0}\right) \quad . \quad . \quad . \quad . \quad (19)$$

The effect of the negative transient, Q_n, is shown in Figs. 9(*a*)–(*c*). At the start of each clamped period C_1 loses a charge Q_n, and thus discharges over a number of cycles to an equilibrium potential which causes the same charge to be replaced during the rest of the cycle.

Since the time-constant R_oC_2 is short, the system reaches a condition of equilibrium by the end of the clamped period, a finite current then flowing through C_1 and R_{in} and corresponding to zero output voltage. If the integrator is to receive no drive during the unclamped period, conditions must not change when the switches open, the same current flow must be maintained, and the input potential must remain zero. The current therefore has a value

$$I_n = \frac{Q_n}{(n+1)\tau} \quad . \quad . \quad . \quad . \quad . \quad . \quad (20)$$

and since it must be supplied during the unclamped periods, it constitutes a current offset. Since the voltage which must be applied is zero, there is no voltage offset.

The magnitudes of Q_p and Q_n are often similar, but they depend very much on the type of transistor and have a wide spread between individuals of the same type. They also depend on temperature, decreasing by about 1% per deg C rise. Q_p is of the order 5×10^{-11} coulomb for audio-frequency transistors and ten or more times smaller for high-frequency types. Thus the combination of a chopping frequency not greater than a few hundred cycles per second and a high-frequency transistor enables the offsets to be reduced to suitably small proportions. For example, if $Q_p = Q_n = 5$ pC, $n = 1$, $\tau = 2{\cdot}5$ millisec, $R_{in} = 300$ ohms and $r_o = 2$ kilohms, then $V_p = 0{\cdot}6\,\mu$V, $I_p = -2{\cdot}3 \times 10^{-3}\,\mu$A and $I_n = +10^{-3}\,\mu$A.

With the transistor held at a stable temperature, the offsets are largely stable, and the drift is negligible.

Various other methods have been proposed for reducing the effects of the transients. Delaying the closure of the output chopper,[2] for example, allows both amplified transients to flow into the integrator, and the net effect is zero if their charges are equal. If the transients are clipped[4] at some stage in the a.c. amplifier, their effects may be reduced, provided that the gain prior to the clipper stage is substantially constant. It has also been suggested that the transients should be ignored, the output signal being observed only during those periods when the transients have died away.

In every case, however, it would appear that the current which flows through C_1 to counterbalance the difference between Q_n and Q_p must be maintained at the same value throughout the cycle if the demodulated output is to be zero. It therefore constitutes a current offset which becomes zero only if $Q_p = Q_n$.

(3.6) A.C. Amplifier

As shown in Fig. 10, the a.c. amplifier is arranged in two sections, each comprising three directly-coupled transistors with overall negative d.c. feedback to stabilize the working point. This arrangement has been described by Chaplin, Candy and Cole[3] and proves both simple and efficient. Good d.c. stabilization can be achieved up to at least 60° C with germanium transistors, provided that J_1 and J_4 have low collector leakage and are operated with a fairly low base current.

The a.c. gain is closely defined by four separate a.c. feedback loops and the coupling resistance R_4, being approximately equal to

$$\frac{R_1}{R_2}\frac{R_3}{R_4}\frac{R_5}{R_6}R_7 \text{ volts/ampere} \quad . \quad . \quad . \quad (21)$$

The input impedance depends on the feedback in the first loop and is chosen to be about 300 ohms. In order to minimize noise, J_1 is operated with a collector voltage below about $0{\cdot}3$ volt, and a collector current of about $100\,\mu$A; consequently J_2 must also have low collector leakage. Transistors of the OC44 or GET874 type can be used for J_1, J_2 and J_4, while the remainder may be, for example, type GET104.

The output stage is arranged to deliver a signal not exceeding

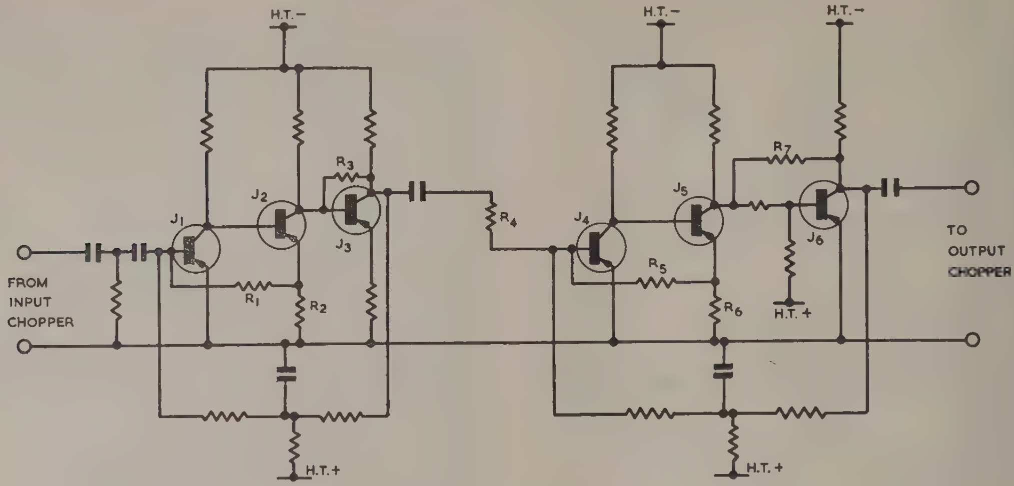


Fig. 10.—A.C. amplifier.

10 volts in amplitude to the output chopper, and to be capable of supplying a peak current of 5 mA at full output voltage.

An important feature is the use of a 2- or 3-stage high-pass filter between the input chopper and the a.c. amplifier. If only a single coupling capacitor is used, variations in the base–emitter voltage of J_1 (due, for example, to changes in ambient temperature) may cause slowly varying currents of substantial magnitude to reach the input chopper, thus giving rise to a varying offset. In a similar manner, d.c. leakage through a single capacitor may cause a relatively steady offset, both effects being proportional to the capacitance and therefore accentuated by the use of low chopping frequencies. Both effects are reduced to negligible proportions by a suitable filter.

(3.7) Integrator

Fig. 11 shows the feedback integrator. Two germanium transistors are employed in an *npn–pnp* configuration which provides cancellation of temperature-induced changes in the base–emitter voltages, together with a phase reversal between input and output. Circuits in which the first stage is a long-tailed pair may give rather less drift, but require at least three transistors and thus present stability problems when the feedback capacitor is connected. Considerably better performance would be obtained by the use of silicon transistors, but the cost would be substantially greater.

The maximum signal applied by the demodulator to the integrator is ± 10 volts, and the maximum current which flows through the integrating capacitor is ± 5 mA. The d.c. amplifier must be capable of supplying this current, and J_2 is therefore operated with a standing collector current of about 5 mA. However, the collector load is kept high, so that the variation in emitter current in the steady state is no more than ± 1 mA and the associated changes in base current and base–emitter voltage are small.

Transistor J_1 is operated with the fairly low collector current of 500 μA in order that changes in gain, with both temperature and time, do not give rise to large changes in base current. The fall in gain and rise in input impedance which occur if the current is further reduced make such reduction not worth while.

Data for the OC140 transistor, with collector currents below 1 mA, are not yet available, and tests were therefore made on nine individuals in order to obtain an approximate estimate of the spread in characteristics. The results are given in Figs. 12(*a*)

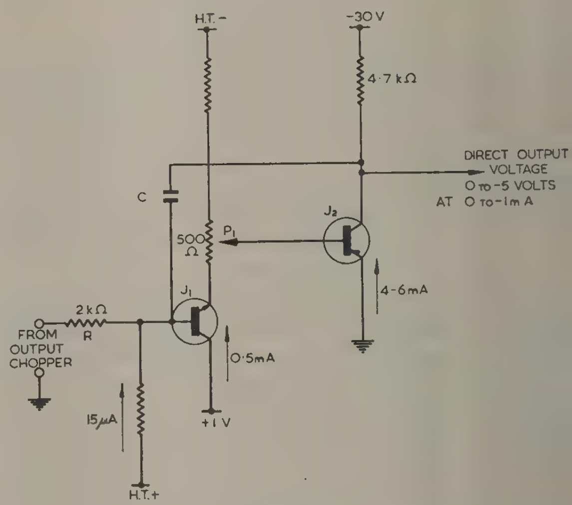


Fig. 11.—Feedback integrator.
J_1: OC140.
J_2: GET113.

and (*b*). Although the number tested is small, one of t transistors had a large signal current gain at 15 mA which w only two-thirds of the manufacturer's minimum figure, and, consequence, the spread can be regarded as pessimistic.

(3.7.1) Initial Adjustment.

It is inferred from Fig. 12(*a*) that the base–emitter voltage J_1, with a collector current of 500 μA, should remain within t limits 100 and 140 mV at 25° C. The spread for J_2 is 17 220 mV, and the difference, which lies between 30 and 120 m is balanced by adjusting the potentiometer P_1 at 25° C with t input terminals joined together, until the output remains stea at about $-2 \cdot 5$ volts.

The nominal base current drawn by J_1 is 15 μA, and a curre of this value is supplied via a fixed resistor to the base of Because of the spread in gain of J_1, however, the base current fact lies within the limits 7 and 23 μA, and a current of ± 8 may therefore flow through the 2-kilohm resistor, causing voltage offset of ± 16 mV, which is compensated by P_1. Wh

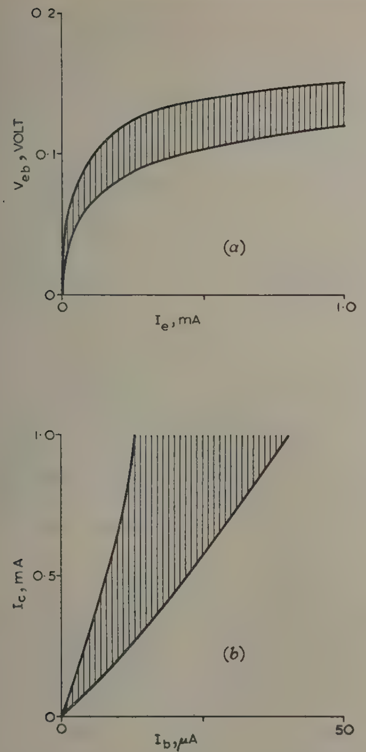


Fig. 12.—Characteristics of OC140 transistor with low collector current, showing spread for nine individuals.

$V_c = -1{\cdot}5$ volts
$T_{amb} = 25°C$

the unit has been adjusted, a current of 8 μA may flow between the input terminals when there is no potential difference between them, but this is of little consequence since the output impedance of the a.c. amplifier and the resistance of the output chopper in the clamped state are both low.

(3.7.2) **Drift due to Temperature Changes.**

The integrator drift voltage is that voltage which must be applied to the input terminals in order to maintain the output at $-2{\cdot}5$ volts. It has several components, and the estimated maximum values of those due to temperature variations of $\pm 25°$ C are listed in Table 1. The current drifts are based on

Table 1

INTEGRATOR VOLTAGE DRIFT WITH TEMPERATURE

Integrator adjusted initially at 25° C.
Output maintained at $-2{\cdot}5$ volts.

	25° C fall	25° C rise
Maximum ΔI_{co} in J_1	$-3\,\mu$A	$+15\,\mu$A
Maximum ΔI_b in J_1	$-3\,\mu$A	$+3\,\mu$A
Total	$-6\,\mu$A	$+18\,\mu$A
Equivalent voltage drift	-12 mV	$+36$ mV
Maximum differential ΔV_{be}	± 10 mV	± 10 mV
Maximum change in terminal voltage	-22 mV	$+46$ mV

As a check on the calculation, 12 integrators were checked for drift between 25 and 50° C, six GET113 transistors being permuted with two OC140 transistors having high and low gains, respectively. The calculated maximum drift in terminal voltage is 46 mV and the measured value in all cases fell between 5 and 17 mV.

the manufacturer's upper limit for collector leakage, and on a temperature coefficient of gain for J_1 of $+0{\cdot}5\%$ per deg C. This figure was closely approximated by all the nine transistors tested. The effect of variations in the collector leakage of J_2 has been neglected since it is relatively small and in a direction which compensates for the effect of leakage in J_1. The figure for differential drift in the base–emitter voltages is based on experience with other transistors and diodes.

(3.7.3) **Ageing Effects.**

The figures in Table 1 are a pessimistic estimate of the short-term drift caused by changes in ambient temperature of $\pm 25°$ C, but they do not allow for ageing effects. These are more difficult to assess accurately and can be compensated by adjusting the control P_1. If re-adjustment is considered undesirable, however, some estimate must be made of the long-term drift, and it is tentatively assumed that ageing effects will not more than double the short-term drift.

(3.7.4) **Gain.**

The forward steady-state voltage gain A_i of the integrator from the input terminals to the output is calculated at 25° C as follows.

A change of $\pm 2{\cdot}5$ volts in the output corresponds to a change of ± 1 mA in the emitter current of J_2, and of about ± 8 mV in the base–emitter voltage. The small-signal current gain is typically about 100, and the base current therefore changes by about $\pm 10\,\mu$A.

That part of P_1 which joins the emitter of J_1 to the base of J_2 has a nominal value of 150 ohms, and the change in the voltage across it is therefore about $\pm 1{\cdot}5$ mV. There is also a small change of about ± 1 mV in the base–emitter voltage of J_1. The small-signal current gain of J_1 is typically 40, the current change in R is $\pm 0{\cdot}25\,\mu$A, and the change in voltage across R is $\pm 0{\cdot}5$ mV. The change in terminal voltage is therefore typically $\pm(8 + 1{\cdot}5 + 1 + 0{\cdot}5) = \pm 11$ mV, and the gain is about 230. Most of the terminal-voltage change is due to the change in base–emitter voltage of J_2, other parameters having only a secondary effect, and it is estimated that the gain should not fall below about 100 in the worst case. Thus the voltage required at the input terminals to swing the output by $\pm 2{\cdot}5$ volts should never exceed about ± 25 mV.

Six measurements, each made using a different GET113 transistor for J_2, gave upper and lower limits of ± 13 and ± 10 mV.

(3.8) Performance

(3.8.1) **Response in Saturated Régime.**

Allowing for the effect of ambient-temperature variations between 0 and 50° C and for a minimum integrator gain of 100, gives $V_t = 46 + 25 \simeq 70$ mV maximum. If the system accuracy is to be $\pm 0{\cdot}1\%$ the output ripple V_{rs} must be several times less than $0{\cdot}1\%$ of 5 volts, e.g. 1 mV. Hence, if $R = 2$ kilohms and $\tau = 2{\cdot}5$ millisec, we have, from eqn. (17),

$$C = \frac{\tau V_t}{R V_{rs}} = 90\,\mu\text{F}$$

Assuming the round figure of 100 μF, the response rate in the saturated condition, given by eqn. (4), is

$$\left(\frac{dV_0}{dt}\right)_{max} = \frac{n}{n+1}\frac{V_{max}}{RC} = 25 \text{ volts/sec}$$

if $n = 1$ and $V_{max} = 10$ volts. This corresponds to a full change of output in $0{\cdot}2$ sec. The increase in V_t which may be expected to arise due to ageing effects causes an increase in the ripple, which, however, remains negligible.

(3.8.2) Dynamic Range.

The a.c. gain required to reduce the current drift I_i, caused by the integrator, to the same level as the current drift of the input chopper ($2 \times 10^{-3}\,\mu$A) is

$$Z = \frac{V_s}{nI_i} = 23 \text{ volts}/\mu\text{A}$$

where $V_s = 46$ mV. A design value of 50 volts/μA allows for a small spread in gain of the a.c. amplifier and for ageing effects in the integrator. The a.c. amplifier and demodulator saturate when the error signal, given by eqn. (5), is

$$I - I_f = \frac{V_{max}}{(n+1)Z} = 0\cdot 1\,\mu\text{A}$$

which is 25 times greater than the smallest detectable signal and implies a useful dynamic range.

(3.8.3) Accuracy and Response Time.

If the feedback resistor is made equal to $1\cdot 25$ megohms, an input of 0–4 μA gives an output of 0–5 volts with an accuracy of $\pm 0\cdot 1\%$. Alternatively, the same output is provided by a input of 0–10 mV if series voltage feedback is employed, wi a feedback factor of 1/500. Larger signals can be amplified wi the same accuracy by increasing the amount of feedback an decreasing the a.c. gain in proportion. Smaller signals are ampl fied with reduced accuracy.

If the input span is 0–4 μA, a step input change of this amou causes the output to change at the limiting rate of 25 volts/s until the error signal is $0\cdot 1\,\mu$A, i.e. $2\cdot 5\%$ of the span, th process occupying $0\cdot 2$ sec. The response then becomes appro mately exponential. The time-constant (5 millisec) given b eqn. (7*a*) is comparable with the cycle duration, and it is ther fore necessary to use eqn. (7*c*), which, in this case, states that th time-constant does not exceed 5 millisec during each unclamp period. Since each period is $2\cdot 5$ millisec, the error is reduced b a factor of $0\cdot 7$ every cycle, and falls from $2\cdot 5$ to $0\cdot 1\%$ in abo 10 cycles. The operation is therefore completed within $0\cdot 25$ se

Figs. 13(*a*) and (*b*) show the typical response to large sine-wa inputs of two different frequencies. Fig. 14 shows a record the amplifier output, when the input is held constant and th ambient temperature is varied from room temperature to 50° As can be seen, the drift is within the calculated limits.

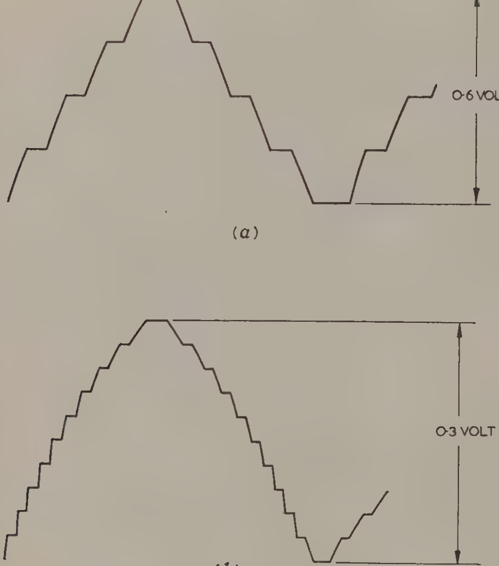


Fig. 13.—Typical output waveforms.
(*a*) Large 25 c/s signal, $C = 50\,\mu$F.
(*b*) Large 10 c/s signal, $C = 250\,\mu$F.
$\tau = 2\cdot 5$ millisec.
$n = 1\cdot 0$.
$V_{max} = 10$ volts.
$R = 1$ kilohm.

(4) CONCLUSIONS

The use of a feedback integrator to smooth the demodulat output of a chopper-type d.c. amplifier enables the whole syste to have a very high forward gain in the steady state and yet b capable of accepting large input signals with a minimum saturation.

Heavy overall negative d.c. feedback can therefore be employ to give a very high and stable net gain together with a fa response at low chopping frequencies.

A practical design has been described which employs a ter perature-stabilized transistor chopper operating at 200 c/s, ar has a drift below $\pm 10\,\mu$V and $\pm 4 \times 10^{-3}\,\mu$A. The gain such that inputs of 0–10 mV or 0–4 μA can be presented as a output of 0–5 volts with an accuracy of $\pm 0\cdot 1\%$.

(5) ACKNOWLEDGMENTS

The authors wish to express their thanks to Messrs. L. Towle and P. S. Boden for useful discussions, and to Mr. D. Harrison for developing the transistor oven. The value References 1–3 is acknowledged, and the Directors of Geor Kent, Ltd., are thanked for giving permission for the paper be published.

(6) REFERENCES

(1) Chaplin, G. B. B., and Owens, A. R.: 'Some Transist Input Stages for High-Gain D.C. Amplifiers', *Proceedin I.E.E.*, Paper No. 2382 M, July, 1957 (**105** B, p. 249).

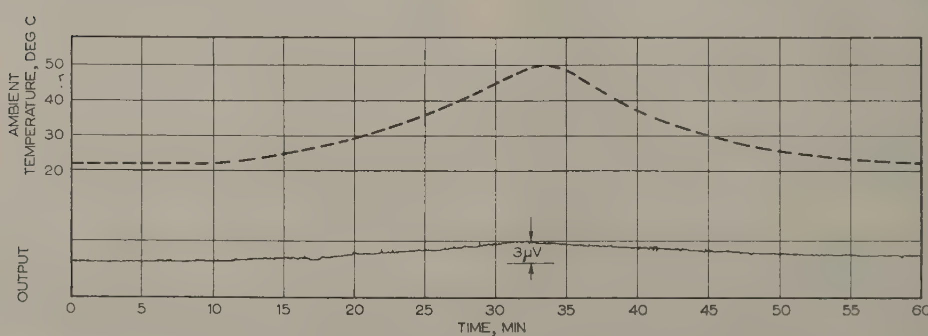


Fig. 14.—Record of output showing drift due to changes in ambient temperature.

(2) CHAPLIN, G. B. B., and OWENS, A. R.: 'A Transistor High-Gain Chopper-Type D.C. Amplifier', *ibid.*, Paper No. 2442 M, November, 1957 (**105** B, p. 258).

(3) CHAPLIN, G. B. B., CANDY, C. J. N., and COLE, A. J.: 'Transistor Stages for Wide-Band Amplifiers', *ibid.*, Paper No. 2892 E, May, 1959 (**106** B, Suppl. 16, p. 762).

(4) HOCHWALD, W., and GERHARD, F. H.: 'A Drift-Compensated Operational D.C. Amplifier Employing a Low-Level Silicon Transistor Chopper', *Proceedings of the National Electronics Conference*, 1958, **14**, p. 798.

(5) HUTCHEON, I. C.: 'The Contact Modulator, Part 3: Modulation Methods', Airpax Electronics Incorporated, Florida, U.S.A., 1959.

(7) APPENDICES

(7.1) Optimum Value of R_s

In the clamped condition the input chopper has a finite resistance r_s and is not able to modulate the input signal efficiently if the source resistance R_s is very low, which may occur when the input signal is a voltage. Since the gain is zero when $R_s = 0$ and ∞, there exists an optimum value of R_s which provides maximum output. It can be shown[5] that the value is

$$R_{s(opt)} = \sqrt{[(n+1)r_sR_{in}]} \quad . \quad . \quad . \quad . \quad (22)$$

and if the actual source resistance is below this value it should be increased by the addition of a suitable series resistance.

(7.2) Circuit with Input Filter

If an input filter is included, as shown in Fig. 15, R_J must be included to prevent the potential of C_J from being held at zero by the short-circuiting action of the switch. If the latter is perfect ($r_s = 0$; $r_o = \infty$) maximum output is obtained[6] when

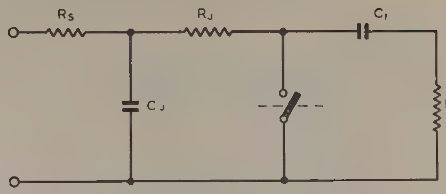


Fig. 15.—Input circuit with filter.

$$R_{J(opt)} = \sqrt{(R_sR_{in})} \quad . \quad . \quad . \quad . \quad . \quad (23)$$

Practical considerations of filter design usually require R_s to be at least a few hundred ohms and, if R_{in} is of the same order, $R_{J(opt)}$ is sufficiently great for the effect of r_s to be neglected. The gain of this circuit is

$$\frac{1}{R_{in} + \dfrac{1}{n+1}\left(\dfrac{R_sR_{in}}{R_J} + R_s + R_J\right)} \quad . \quad . \quad . \quad (24)$$

which may be compared with the higher value

$$\frac{1}{R_{in} + \dfrac{R_s}{n+1}} \quad . \quad . \quad . \quad . \quad . \quad . \quad (25)$$

which is provided by the same circuit without the filter components C_J and R_J.

DISCUSSION BEFORE THE MEASUREMENT AND CONTROL SECTION, 15TH MARCH, 1960

Dr. G. B. B. Chaplin: Much has been published in the past concerning methods of reducing drift in transistors, but the instrument designer is still left with many problems. He has to choose the best system for his particular application and he has also to solve many engineering problems, such as ensuring that the drift is not critically dependent on power supplies, and that the demodulator can give sufficient power to provide rapid response.

The paper presents an interesting solution to these various problems and describes a practical amplifier with drifts of only a few microvolts and a few millimicroamperes. The authors have made two main decisions. The first is to control the temperature of the input chopper, and the second is to provide an output integrator with plenty of gain.

The reason for using temperature control is presumably to keep the voltage drift below about 20μV over a 60° C temperature range. Although a silicon transistor would reduce the current drift to a negligible value without resorting to temperature control, it would not significantly affect the voltage drift.

The temperature-controlled oven and its associated controlling amplifier significantly increase the complexity of the d.c. amplifier, and the system can be justified only if there is no reasonable alternative.

A particularly attractive solution would be for the transistor manufacturers to select pairs of chopper transistors in which the voltage drifts were matched to within 10%. Provided that the junction temperatures are kept within 2 or 3° C of each other, the differential drift should not be greater than 20μV over the 60° C range of temperature.

With regard to the output integrator, I feel this is a very elegant solution to the demodulation problem, although the particular circuit used (Fig. 11) has a relatively large drift of 48 mV and also requires adjustment. A simple long-tailed pair would have a much lower voltage drift and would not require adjustment. The authors have rejected the long-tailed pair on the grounds that there would be an extra transistor in the feedback loop, but I feel that this argument is not strictly correct. The two transistors are not in cascade and therefore should not contribute a double phase shift.

Mr. K. P. P. Nambiar: I would like to refer to some properties of the silicon alloy transistor in relation to its application in the input chopper stage. Measurements carried out on a number of these units show that, at 0·5 mA base current, the drift in saturation potential ranges from 100 to 250 μV when the ambient temperature is raised from 20 to 85° C. At an optimum base current of 1 mA, however, the potential drift was less than 50 μV in 80% of the units tested. In a batch of silicon alloy transistors, the lowest output impedance in the off state of the silicon alloy transistor, under circuit conditions of zero leakage current, was found to be of the order of 600 kilohms at 20° C dropping to about 320 kilohms at 100° C. Taking 300 kilohms as the lowest impedance with any unit at 85° C, the current drift resulting from the variation in saturation potential and finite output impedance of the transistor is given by $1{\cdot}6 \times 10^{-4}\,\mu$A for the range 20–85° C. Thus if the potential drift and the off-state impedance are the important factors, the choice of silicon alloy transistor in the input stage would appear to be an appropriate solution.

A technique for reducing the on-state potential drift with temperature, which we have recently tried out, is illustrated in Fig. A. A base switching current much lower than 1 mA is assumed. The drift compensation results from the increase in resistance of the silicon resistor R_x, known as the Sensistor,

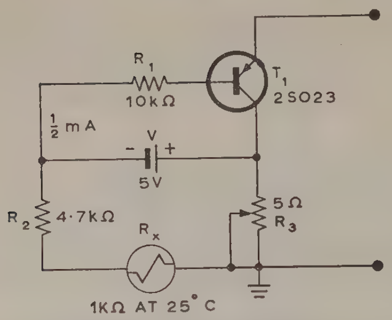


Fig. A.—Compensation for the potential drifts with temperature by silicon Sensistor resistor.

with temperature. The Sensistor resistance increases by about 50% when the temperature is raised from 20 to 85° C, so that a 10% drop in the saturation potential of the transistor will be adequately compensated by the values of R_2 and R_x shown in Fig. A. The drift in potential could be reduced to within 50 μV in the temperature range of 20–85° C using this technique; by

tion of the switching waveform. The clamping period may b adjusted by the choice of the potential V_r or the capacitance

It would thus appear that, by keeping the output chopper the on state during the periods when the transients occur, it possible altogether to suppress the effects of transients.

Dr. A. J. Maddock: The authors state that their circuit is the nature of a drift-corrected amplifier. When one is usi switching contacts of the mechanical type it is more usual arrange these in change-over form so that, during the correcti period, the input is disconnected and there is direct connectio between output and input of the amplifier giving 100% insta taneous feedback, instead of passing through the rather hi feedback resistance as shown. Why did the authors not u that system? It may be because, with the transistor switch, o does not get truly open-circuit conditions to isolate the feedba connection.

I agree with the general level of zero stability of about 1 μV f the switching circuit when using commercial types of mechanic chopper (vibrating reed). The authors say that the figure 10 μV in their circuit. Dr. Chaplin indicates some ways which this might be improved. I would like to ask the author if they had to improve the stability, in what direction wou

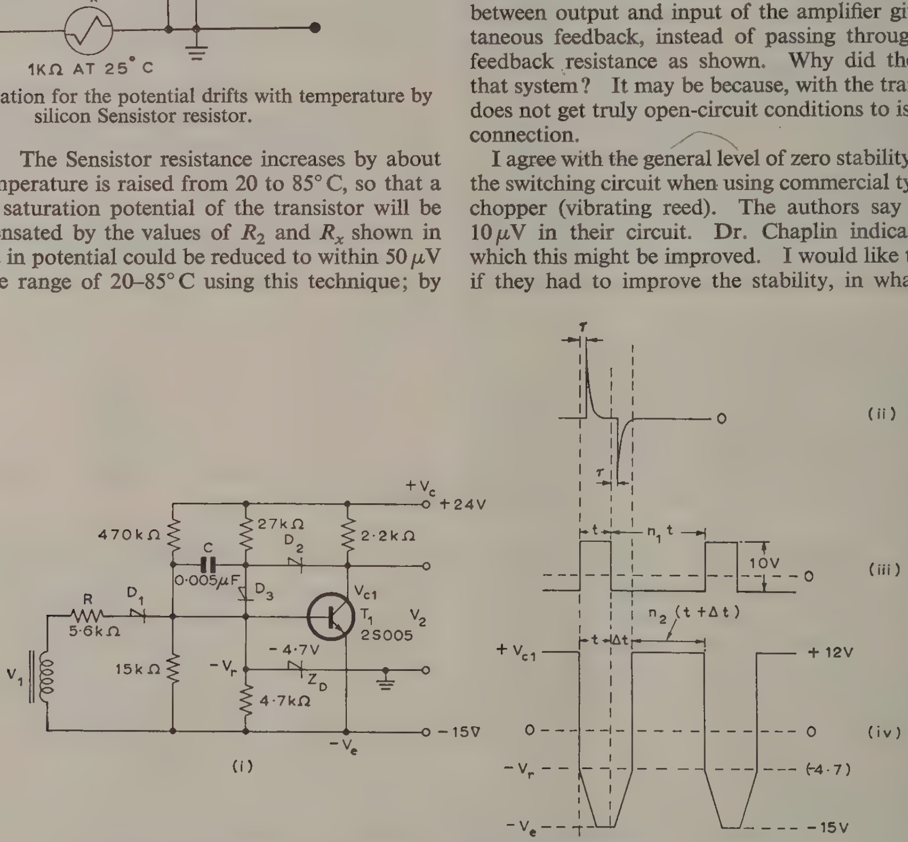


Fig. B.—Expansion of the clamping period for the suppression of transients.

(i) Circuit.
(ii) Transients appearing at the output of the a.c. amplifier after the amplifier time delay of τ seconds.
(iii) Input V_1 from the chopper waveform generator.
(iv) Output V_2 at the collector of T_1 for application to the demodulator.

careful adjustment of R_2 it is also possible to obtain a drift of $\pm 10\,\mu$V over a temperature range of 25–50° C.

The oven temperature of 50° C mentioned in the paper appears to be rather low, particularly for equipments which may be used in tropical climates where crystal ovens in telecommunication equipments are invariably maintained at 75° C or above.

My last point concerns the transients. It appears that the transient occurring at the instant of switching on the input chopper transistor can be clamped at the output by the simultaneous application of the demodulator waveform to the base of the output chopper transistor. The transient occurring at the instant of switching off can be suppressed by extending the period of clamping of the demodulator. The extension of the clamping period may be achieved by the use of a gated Miller circuit shown in Fig. B, which is a variation of that mentioned by Chaplin and Owens[2] and designed to integrate the lagging por-

they go, what are the limitations giving rise to their 10 μV, an how might they attack this problem?

Mr. E. H. Cooke-Yarborough: I would like to draw attentio to an arrangement* which is perhaps halfway between th straightforward directly coupled transistor amplifier and th chopper type.

Fig. C shows the common-emitter input circuit of the d. amplifier. The base current is made up of the two component shown. Any variation in the sum of these appears as curre drift. In recent silicon transistors I_{c0} is of the order of $10^{-2}\,\mu$A and so the component $I_c(1-\alpha)/\alpha$ is the dominant one. Effor to reduce this component by reducing I_c fail when the emitte current is reduced below about 10 μA, because, at lower emitte

* Cooke-Yarborough, E. H.: 'The Pulsed Direct Current Amplifier: A Metho of Operating Transistors at very Low Currents', A.E.R.E. Memorandum No. 47 Patent Application No. P.37337/59, 3rd November, 1959.

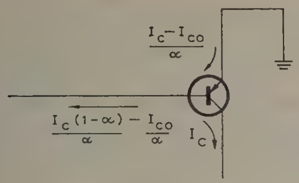


Fig. C

currents, the carrier injection efficiency falls off, reducing the current gain, α.

This difficulty can be avoided by causing the transistor to conduct in short pulses, the input current being stored in a capacitor in the intervals between pulses. One way of doing this is shown in Fig. D, short negative pulses being applied at

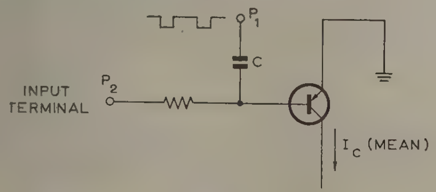


Fig. D

P_1 while the input current to be measured flows out at P_2. For a given peak emitter current the mean base and collector currents are reduced by a factor equal to the ratio between pulse length and pulse spacing.

It can be shown that the pulse length need only be of the order of the mean carrier transit time across the base of the transistor. Thus, if the transistor has an α cut-off frequency of 20 Mc/s, the pulse length may be 10 millimicrosec, and if the pulse rate is 10^6 pulses/sec, a hundredfold reduction in base current can result.

This arrangement has no advantage over the conventional direct-coupled amplifier so far as voltage drift and the effect of I_{c0} are concerned, but it has the advantages over the normal chopper amplifier in that a faster response is obtainable and no demodulator is needed, since only the d.c. component of the collector current need be amplified by later stages.

Mr. W. I. McMillan: We have been developing an instrument similar to the one described, but we decided to keep the sampling frequency high in order that the amplifier performance would not be limited by transistor noise.

In order to compensate for the increased voltage drift due to switching transients we chose to use an oven with a temperature error of $+0\cdot5°$ C. This brings other compensations if the amplifier is being used with thermocouples as the input transistor oven can be used as a 'cold' junction, and in some cases this is very economical. In most amplifiers of this type a potentiometer chain is used to derive a voltage to balance the emitter-collector voltage of the switch when closed. In order to keep the amplifier drift below 1 μV over a large temperature range it is necessary to keep the difference in temperature coefficient of the two parts of the potentiometer below 25×10^{-6}. This can be a real limit to the performance of an industrial amplifier.

Mr. E. P. Fowler: We have also been developing a very similar amplifier but have arrived at different results. As we were measuring current in an ionization chamber we could tolerate 100 μV drift, and thus we never considered an input chopper oven. We did, however, want to reduce the current drift, and so we used a silicon-transistor chopper. For the output circuit we used a long-tailed-pair first stage followed by a transistor having a feedback resistor connected to its emitter and the load

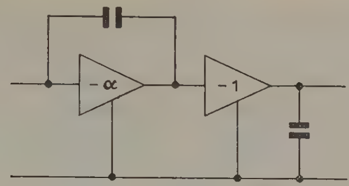


Fig. E

to its collector. This is shown schematically in Fig. E; the system has an α controlled to within 10% and no overall phase reversal.

At very high frequencies the gain can be reduced very nearly to unity and the integrating effect of the capacitor will be very small. For that reason, we connected another across the output to earth.

With the authors' amplifier I presume that at high frequencies, if the gain becomes zero, there will be no phase reversal and the whole amplifier may get unstable.

By having more gain in the output amplifier we have a response of 17 millisec for the step function from zero to the full output of the amplifier instead of the large fraction of a second quoted in the paper. By having lower drift we do not need so much gain in the main a.c. amplifier, and using only three transistors with an open-loop gain of 1 kV/μA we will have an accuracy of $0\cdot1$% for 0–10 μA input.

In Fig. 11, the authors refer to 'h.t.—' and '—30 volts'. Which is the higher?

Mr. R. B. Stephens: It has been suggested that, because the thermocouples are relatively slow-acting devices, an amplifier having a frequency response extending from zero to a few tens of cycles is adequate. In fact, there are applications requiring a much wider bandwidth, for example in large data-handling systems in which common amplifiers and digitizing circuits are used to measure the voltages arising from a large number of thermocouples or strain gauges which are scanned sequentially.

Usually a measuring accuracy within $\pm 0\cdot1$% is specified and the time available for each measurement may be only tens of milliseconds or less. In these cases, an amplifier is required having a drift of the order of 10 μV and a bandwidth of upwards of 1 kc/s for $0\cdot1$% accuracy.

I should like to ask the authors whether their relatively narrow-band amplifier could be used to correct the drift of a wide-band d.c. amplifier, so that the combination would have all the desirable features of the d.c. amplifier described and, in addition, a wide bandwidth.

Mr. P. M. Thompson: Mr. Cooke-Yarborough brought his pulsed d.c. amplifier to my attention in May, 1959. It is possible to extend his technique to average input currents much lower

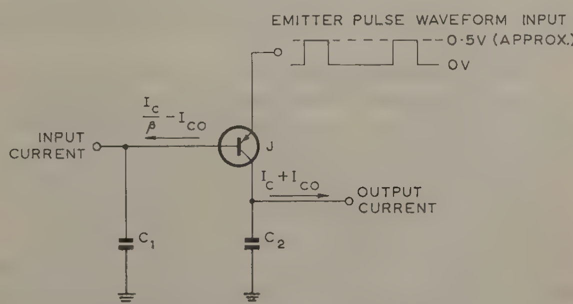


Fig. F.—Pulsed current amplifier.

than the normal leakage current of a silicon transistor by arranging that the transistor has no bias voltage across either junction during the quiescent period between current pulses. A circuit by which this may be achieved is shown in Fig. F. It differs from Mr. Cooke-Yarborough's circuit in that the input storage capacitor C_1 is earthed, and the pulse is applied in the opposite sense to the emitter. The silicon transistor is operated with zero collector-base voltage. The emitter also is held at zero voltage during the quiescent period, and is raised to the correct potential to cause current flow when pulsed. Th minimum average input current is limited by the total bas current which flows during the quiescent period. A practic limit is about $10^{-5}\,\mu$A, which is achieved with emitter-base an collector-base voltages of approximately 20 mV. At this inpu current, a mark/space ratio of 1 : 10^5 would be required to rai the peak base current to 1 μA. Amplifier stages of this typ may be readily connected in cascade, as the potentials of th input and output terminals are similar.

THE AUTHORS' REPLY TO THE ABOVE DISCUSSION

Messrs. I. C. Hutcheon and **D. Summers** (*in reply*): As Dr. Chaplin points out, the design of an amplifier system is closely associated with the design of its power supplies, and with the methods used to solve other problems of a practical nature. A block schematic of the whole system is therefore given in Fig. G.

Advantages of this arrangement include the ability to operate silicon transistor in a simple chopping circuit requiring n coupling transformer. If the Zener diode associated with th input chopper is also mounted in the oven, it can be an inexpe sive type having a breakdown voltage in the 6–7-volt region an a low slope resistance. The component cost of the oven an its associated amplifier is below £5, and the area required on printed circuit board is about 5 in^2.

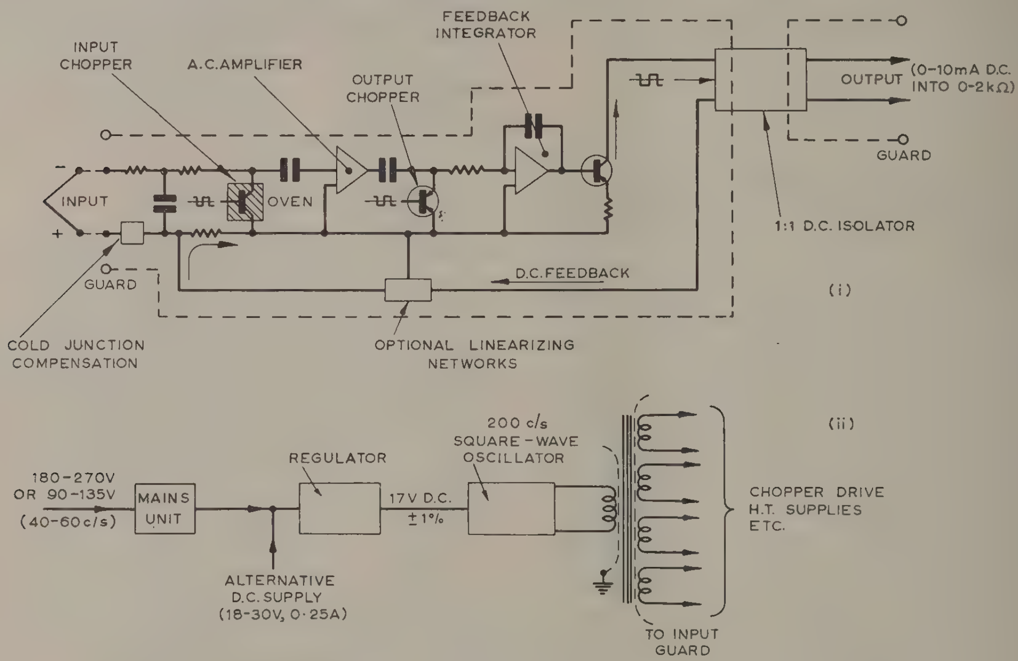


Fig. G.—Complete amplifier system.

(i) Amplifier and d.c. isolator. (ii) Power supply arrangement.

from a wide range of a.c. and d.c. supplies, and to provide almost infinite rejection of common-mode signals up to several hundred volts, a.c. or d.c.

A possible disadvantage is the limited frequency response of the d.c. isolating stage. Biased-diode networks may be incorporated in the feedback to correct for non-linear thermocouple characteristics—up to 25 volts being available for this purpose.

We agree with Dr. Chaplin that the provision by transistor manufacturers of matched pairs of chopper transistors would be very useful. High-frequency silicon transistors would be the most suitable if they could be provided at reasonable cost.

Stabilizing the temperature of the chopper transistor has several advantages, however. It stabilizes both the V_{ce} of the transistor and the offset caused by chopper transients, and so allows the use of a single relatively inexpensive low-frequency

In reply to Dr. Maddock, the major residual cause of volta drift is a hysteresis effect in the transistor, whose offset is fou to change by several microvolts during the first few hours aft a step change in temperature. If this effect is eliminated by preliminary soak, the whole system is stable within 2 or 3 μ during test runs of 100 hours which include ambient te perature extremes of 20° and 50° C. These results are almo certainly better than could be obtained without temperatu stabilization. The hysteresis effect is not fully understood a merits investigation by the manufacturers of transistors intend for use as choppers.

We agree with Mr. Nambiar that the chopper transistor shou be silicon, but feel that adjusting the temperature compensati provided by a Sensistor might prove a somewhat expensi procedure in production. The oven temperature has now be

raised to 60° C, which corresponds to the upper temperature limit for the rest of the system when constructed with germanium transistors, and this is thought to be adequate for most industrial requirements.

Mr. Nambiar's proposal to clamp out both transients would appear to be a complete solution only if the transients were of equal area. If they are not, a charge accumulates on the input coupling capacitor, and a current flows in the input circuit while the output chopper is clamped. If there is to be no output during the unclamped period, this current must be maintained and so constitutes an offset.

We agree with Dr. Maddock that transistor switches would be inadequate for the method of drift correction which he describes. In the system adopted, the loop gain exceeds 2000 and this is sufficient.

Both Dr. Chaplin and Mr. Fowler have suggested that the first stage of the integrator should be a long-tailed pair. Our allowance of 48 mV drift includes a large safety factor, and the worst measured figure, including the effect of current drift acting through the 2-kilohm integrating resistor, was 12 mV. We doubt whether a germanium long-tailed pair would give much better results. The gain of our circuit (nominally 200) is such that an input swing of ± 12 mV is required to swing the output fully, and in this respect a long-tailed-pair circuit should give some improvement. We agree that the need for an initial adjustment would be less in a circuit using two identical transistors with low collector currents.

Several circuits were constructed using long-tailed pairs, and found to be prone to instability, but our explanation may not have been the correct one.

The method of obtaining low current drift described by Mr. Cooke-Yarborough and its extension by Mr. Thompson are most ingenious, although unfortunately not applicable to the measurement of low voltages.

Mr. McMillan's use of the oven to stabilize the cold junction of a thermocouple represents a useful economy, but we have preferred to avoid the practical problems of carrying wires of thermocouple material into the amplifier, and of achieving the necessarily higher degree of temperature stabilization.

We have not as yet found it necessary to connect a capacitor across the output of the integrator, but Mr. Fowler's argument on this point appears to be valid. The rather long response time (0·25 sec) which we quoted is a function of the integrator performance and would be reduced to about 80 millisec if no safety factor were allowed. A further reduction is obtainable if more than 0·1% ripple is tolerable in the output. The small-signal bandwidth is, of course, higher than the large-signal response would suggest, and a figure between 15 and 30 c/s is typical for the amplifier less the d.c. isolator.

In the reference to 'h.t.—' the voltage is immaterial provided that the specified current is obtained.

In reply to Mr. Stevens, we see no reason why the amplifier described should not be used to correct the drift of a wide-band d.c. amplifier, and would refer him to Dr. Chaplin's earlier paper in which this question was examined. There is, however, no necessity to use a wide-band d.c. amplifier in thermocouple scanning systems. A simple a.c. amplifier is equally satisfactory, provided that the scanning switches are arranged to clamp both input and output to earth at regular intervals* so as to reset the coupling-capacitor potentials and correct for drift.

* HUTCHEON, I. C.: 'An Iterative Analogue Computer for Use with Resistance Network Analogues', *British Journal of Applied Physics*, 1957, **8**, p. 370.

621.317.39

The Institution of Electrical Enginee
Paper No. 3256
Sept. 19
(

AN AUTOMATIC CONTINUOUS DENIER RECORDER FOR SYNTHETIC TEXTILE YARNS

By C. D. RUTTER, B.Sc.(Eng.), R. D. WRIGHT, B.Sc.(Eng.), Associate Member, R. N. ALDRICH-SMITH M.Sc., F.Inst.P., and E. J. R. HEWITT, M.A., Associate Member.

(*The paper was first received* 20*th November*, 1959, *and in revised form* 25*th February*, 1960.)

SUMMARY

Measuring the diameter of a fine textile filament by its electrical capacitance effect between the fixed plates of a capacitor presents the twin problems of the very small capacitance change to be measured and the high degree of stability to be maintained in the capacitor.

The present instrument is insensitive to changes in the standing capacitance and achieves this by vibrating the yarn in and out of the gap. The capacitance between the two electrodes is therefore varying at the frequency of the vibrating yarn, and a variable capacitor is switched into circuit in phase with the yarn and used to balance this change in capacitance in a bridge circuit. The output from the bridge is amplified and used to drive a motor which balances the bridge by adjusting the variable capacitor. The movement of the motor also drives a pen recorder. The bridge circuit conveniently provides several measuring ranges and allows the measuring head to be one of several connected to the main instrument by long cables.

Synthetic-fibre manufacture involves numerous identical production units in parallel. Monitoring and control gear is therefore fed from a multiplicity of transducers which must be matched, reliable and capable of simple routine test and replacement without upsetting their related circuits.

(1) INTRODUCTION

Continuous filament yarns of synthetic fibres such as nylon are produced by extruding molten polymer through a small hole in a die. It cools in air, solidifies and is wound up on bobbins. Factors such as extrusion rate and wind-up speed may cause undesirable variations in the line density of the finished yarn, and it is essential to make measurements of yarn line density at regular intervals, or preferably continuously. The thickness of a yarn is specified, for a range of materials including nylon, in terms of denier, which is the weight in grammes of a 9 000 m length of yarn. To monitor the denier it is necessary to measure the mass per unit length of the yarn, and the standard method is to reel off a measured length, usually 90 m, weigh it and convert this to denier. This is a time-consuming operation only possible at widely spaced intervals, and since it determines the mean denier over a fairly long length of yarn, ignoring short-term variations, is of limited use for quality control.

One method of continuously monitoring a moving thread line is to make use of the increase in capacitance caused by the insertion of yarn between two parallel-plate electrodes. This capacitance change has been investigated theoretically by Mack,[1, 2, 3, 4] the latter reference containing a summary of the results. For a cylindrical filament the capacitance change varies with the square of the filament radius and inversely as the square of the electrode separation:

$$\Delta C = 2\epsilon_0 \frac{\epsilon - 1}{\epsilon + 1} \frac{r^2}{D^2} \quad . \quad . \quad . \quad . \quad . \quad (1)$$

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.
The authors are with British Nylon Spinners, Ltd.

Where ΔC = Capacitance change per unit length, F/m.
ϵ = Dielectric constant of the material.
r = Filament radius, m.
D = Electrode separation, m.
ϵ_0 = 8·854 pF/m.

Thus, if the density and dielectric constant remain constant alo its length, the total capacitance change is proportional to t mass of yarn between the electrodes. The density of nylon such that, for a solid cylindrical filament, the denier N is giv by $N = 20{\cdot}8r^2$, where r is measured in mils. The capacitan change per inch of electrode length, assuming the dielectr constant of nylon to be 4, is therefore given by

$$\Delta C = 40{\cdot}8 \frac{N}{d^2} \times 10^{-3} \text{ picofarad}$$

where d, the electrode separation, is measured in mils.

In eqn. (1) it is assumed that the diameter of the yarn is sma compared with the separation of the electrodes, and that the ya is not near either electrode. When these conditions are n fulfilled, the capacitance change is greater than that indicated the formula by an amount which increases with increasi dielectric constant. The error obtained experimentally is 7 for a cylinder of dielectric constant 3 centrally placed betwe electrodes separated by 2½ times the filament diameter. Th error increases to 20% when the filament is touching one ele trode. Wide electrode spacing improves linearity, but gives on small capacitance changes.

Most textile yarns consist of a number of separate filamen and not the single cylindrical filament (monofil) used in t derivation of eqn. (1). However, if the yarn is sufficient tightly twisted the individual filaments are held together to for a yarn of cylindrical cross-section which can be measured wi capacitance instruments. If it is desired to test yarn with sma or negligible twist, false twist may be introduced locally into t yarn passing between the electrodes.

(2) DESIGN CONSIDERATIONS

(2.1) Degree of Stability

This principle of measuring the capacitance has been used a number of instruments.[5, 6, 7, 8] However, since the capacitan changes caused by inserting the yarn in a gap of practical dime sion are very small compared with the standing capacitance these instruments are unsuitable for long-term continuo measurement of denier. For example, if we insert a 15-deni nylon yarn, consisting of a single filament 0·001 7 in in diamete between electrodes 2 in long, the capacitance increase is calculate to be 0·049 pF when the electrodes are separated by 0·005 i and 0·003 pF when the electrode separation is 0·020 in. Bo these examples neglect the error due to the narrow-gap no linearity. The standing inter-electrode capacitances when t electrodes are 0·1 in wide are about 10 and 3 pF, respectivel

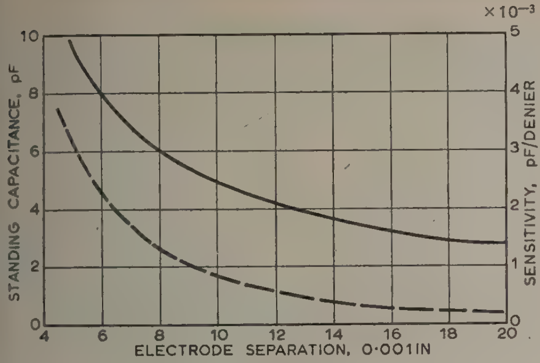


Fig. 1.—Standing capacitance and increase in capacitance due to insertion of yarn versus electrode separation.

Electrode dimensions
Length: 2 in.
Width: 0·1 in.
—— Standing capacitance.
---- Sensitivity.

Figures are reproduced in Fig. 1 for standing inter-electrode capacitances and for increases in capacitance due to yarn/electrode separation. It will be seen that the percentage change in capacitance due to the insertion of yarn is very small, and any small increase or decrease in standing capacitance will have a large effect on the instrument. The stability demanded of the standing capacitance is of the order of 10 parts per million and would be impossible to achieve under normal plant conditions, although possible under good laboratory conditions. For this reason the existing instruments are unsuitable for long-term continuous measurement of denier, although they may be used for measuring and recording changes in denier occurring in short lengths of yarn up to a few metres. Measurement of the actual denier at any point in the yarn can be made only by setting the instrument to zero without yarn, inserting the yarn between the electrodes and reading the change in output deflection. This method is used to calibrate the instrument against a standard yarn.

(2.2) Moisture Effect

Eqn. (1) is modified by the effect of absorbed moisture and by differing packing density of the individual filaments. In practice, the most troublesome change is that due to varying quantities of absorbed moisture.[9, 10] This effect can be reduced by making the capacitance measurements at a high frequency, and measurements on the thickness of single fibres have been made at frequencies[11] up to 26 Gc/s. Useful results can be obtained down to about 500 kc/s, but lower frequencies not not suitable for measurement purposes because of electrostatically produced noise.

(2.3) Temperature Effect

The temperature coefficient of the dielectric constant of nylon is a factor giving rise to variations small enough under production conditions to justify merely a calibrated compensating potentiometer in the measuring circuit, this being set according to the reading of a small thermometer located at the detecting head.

(2.4) Electrostatic Noise

Tests at 3 kc/s showed that noise was very troublesome with monofil yarn, and with low-twist multi-filament yarn the signal was completely masked by noise. This noise effect disappeared when cotton was used instead of nylon.

The noise is due to the generation of electrostatic charges when the yarn passes over the guides, and is not dependent on the supplies to the measuring device, showing that the effect was produced by the amplification of noise and not by capacitance changes. The noise was not reduced by the presence of ionized air produced by a spark discharge or a β-ray source. The magnitude of the noise made the operation of the bridge at low frequencies quite impossible, at any rate for measurements on nylon yarns. The noise was found to be reduced at higher frequencies, and at 1 Mc/s it was less than 10 μV in a bandwidth of 10 kc/s. It was decided to use 1 Mc/s because it is the highest frequency suitable for normal bridge circuits.

(2.5) Bridge Methods

Methods of measuring small capacitance changes in existing instruments[5, 6, 7, 8] suffer from the disadvantage of having an electronic valve as part of the capacitance detecting circuit. Random capacitance changes in the valve will appear to the circuit as changes in yarn denier and give rise to a background noise level which limits the sensitivity of the circuit. Improved performance can be obtained from bridge circuits involving only passive elements, and a single bridge circuit is already in use in a commercially available evenness recording equipment.[7]

Besides the greater stability that can be obtained, suitable bridge circuits have further advantages. The yarn electrodes can be connected as a 3-terminal capacitor with the bridge arranged to reject earth capacitances. Also capacitance changes of a convenient magnitude may be made in one part of the circuit to balance the very small capacitance changes between the electrodes, and only the electrodes need to be made in an extremely rigid manner as conventional components may be used in the rest of the circuit provided that the bridge elements are completely screened.

Bridge circuits with one or more inductively coupled ratio arms have been used for measuring small capacitances of the order of those encountered in this type of instrument; these circuits were first described by Blumlein,[12] and the simplest forms are shown in Fig. 2. The great advantage of such circuits

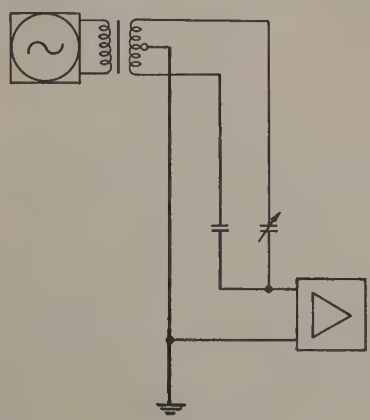


Fig. 2.—Form of bridge circuit with one or more inductively coupled ratio arms.

is that the effect of stray capacitances can be eliminated without the completion of a Wagner earth. This feature has been thoroughly discussed by Clark and Vanderlyn.[13] The circuit-elements should be thoroughly screened and the screening taken to the transformer centre tap, which is conveniently earthed.

Also, there should be a low leakage inductance between the two halves of the winding and a low winding resistance.

By using transformers with winding ratios other than 1 : 1, bridges with different ratios between the measured and balancing elements may be constructed. Even greater flexibility in circuit arrangement may be obtained by using two transformers, each with a number of taps.[13] Circuits working on these principles have been employed to measure very small capacitance changes.[14, 15, 16] For yarn evenness measurements, however, it is essential to work at frequencies above 500 kc/s to overcome electrostatic noise, and although similar bridges have been operated at radio frequencies,[17] it is difficult to design transformers for radio-frequency operation with a low leakage inductance, especially when a high turns ratio is required. In practice, it is convenient to use a 1 : 1 transformer.

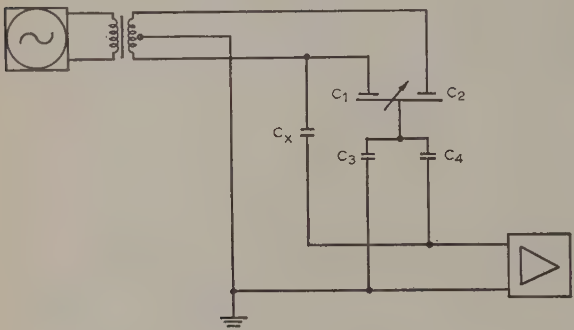


Fig. 3.—A convenient bridge circuit.

A more convenient circuit is shown in Fig. 3, and instruments designed for measuring valve inter-electrode capacitances are commercially available using this circuit.[18, 19, 20, 21] The advantage of this circuit is that a large ratio may be obtained between the variable balance capacitor and the measured capacitance. The output voltages on the two halves of the transformer secondary remain equal and 180° out of phase if the leakage reactance between the two halves is sufficiently low and the winding resistance is low. The equation for balance is

$$Y_x = \frac{Y_4(Y_2 - Y_1)}{(Y_1 + Y_2 + Y_3 + Y_4)} \quad . \quad . \quad . \quad . \quad (2)$$

and assuming that all the circuit-elements are lossless capacitors as in the circuit of Fig. 3 the balance equation is

$$C_x = \frac{C_4(C_2 - C_1)}{(C_1 + C_2 + C_3 + C_4)} \quad . \quad . \quad . \quad . \quad (3)$$

C_1 and C_2 may be the two parts of a differential capacitor, and in this case $(C_1 + C_2)$ is constant and the rotation of the variable capacitor is directly proportional to the unknown capacitance. The constant for Fig. 3 is then $C_4/(C_1 + C_2 + C_3 + C_4)$. This ratio can be made as small as is desired by increasing C_3 or decreasing C_4. When it is desired to make the ratio very small C_4 may become an inconveniently small capacitance, and in this case a T-network may be used in place of C_4. The circuit then becomes that of Fig. 4, the balance conditions being given by

$$C_x = \alpha\beta(C_2 - C_1) \quad . \quad . \quad . \quad . \quad . \quad (4)$$

where $\alpha = C_4/(C_1 + C_2 + C_3 + C_4)$

and $\beta = C_8/[(1 - \alpha)C_4 + C_7 + C_8]$

If additional balance controls are required, networks may be

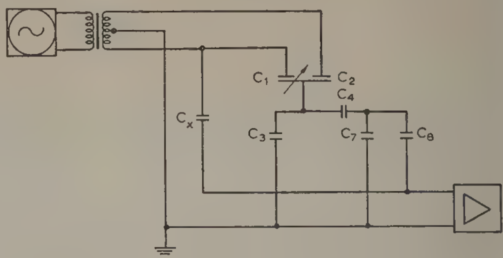


Fig. 4.—Circuit of Fig. 3 modified when the ratio $C_4/(C_1 + C_2 + C_3 + C_4)$ is very small.

connected in parallel as shown in Fig. 5, and the balance equatio
for the circuit becomes

$$Y_x = \frac{Y_4(Y_2 - Y_1)}{(Y_1 + Y_2 + Y_3 + Y_4)} + \frac{Y_4'(Y_2' - Y_1')}{(Y_1' + Y_2' + Y_3' + Y_4')} \quad (5$$

One of the two networks may be resistive, as in Fig. 5, so tha
the conductive component of the unknown admittance can b
balanced. In this circuit the unknown conductance is nc

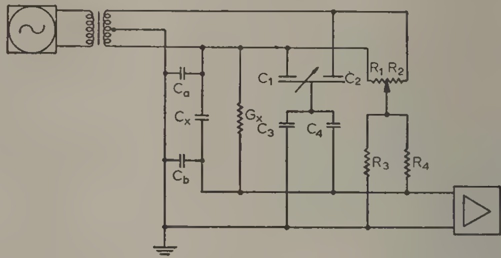


Fig. 5.—Bridge circuit modified to provide additional balance controls.

linearly related to the potentiometer rotation, since, whil
$(R_1 + R_2)$ is constant, $(1/R_1 + 1/R_2)$ is not, but the effect c
this can be reduced by connecting fixed resistors in series with th
sliders of the potentiometer so that the effective change i
$1/R_1 + 1/R_2$ is small.

If two different balance controls are required it is sometime
convenient to connect the two networks in series as in Fig. (
the balance equation then being

$$Y_x = \beta[(Y_6 - Y_5) + \alpha(Y_2 - Y_1)] \quad . \quad . \quad . \quad (6$$

where $\alpha = Y_4/(Y_1 + Y_2 + Y_3 + Y_4)$

and $\beta = Y_8/[Y_4(1 - \alpha) + Y_5 + Y_6 + Y_7 + Y_8]$

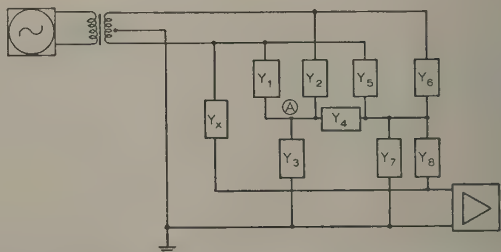


Fig. 6.—Bridge circuit with two networks connected in series, thus providing two different balance controls.

Vhen all the circuit-elements are lossless capacitors Y_2, Y_1 and Y_5, Y_6 may be the admittances of differential variable capacitors, nd we are left with α and β as constant range factors.

This type of network could be extended to include further alance controls, thus obtaining further balance conditions of he following form:

$$Y_x = \gamma\{Y_9 - Y_{10} + \beta[Y_6 - Y_5 + \alpha(Y_2 - Y_1)]\} \quad . \quad (7)$$

In all these circuits Y_x represents the admittance between the wo electrodes through which the yarn is passing. These elec-rodes are mounted together in a solidly constructed and care-ully screened measuring head, which is connected to the rest of he bridge circuit by screened leads. C_a and C_b (in Fig. 5) epresent capacitances to earth and include the lead capacitances, C_a appears across one half of the transformer secondary winding, ut because the two halves are tightly coupled, the voltages on he terminals remain equal and opposite in phase despite the urrent drawn through C_1, and so the balance of the bridge is naffected. C_b appears across the input to the amplifier, and o may cause a change in the output voltage but not in the ridge balance. In order to obtain increased sensitivity the nput to the amplifier may be tuned, and the amplifier input mpedance may then be quite high even though C_b may be large wing to long cables joining the head to the bridge. The tuned ircuit will need adjustment if the length of cable to the head s changed.

The bridge output voltage can be used as a measure of the apacitance change when only yarn evenness measurements are equired, but the calibration is not necessarily constant, since he output reading is dependent on the oscillator voltage and he amplifier gain, both of which may vary. These difficulties f stability in the measuring head and gain in the oscillator and mplifier circuits can be overcome if the value of ΔC, due to the arn, can be converted to a phase-modulated signal, which can hen be rectified by a phase detector, discriminating against any amplitude-modulated or d.c. signal due to other parts of the circuit. The detected phase-modulated signal can then be amplified and used to operate a servo motor driving a self-balancing bridge balancing the unknown capacitance change. The phase-modulated yarn signal can be produced if the yarn is regularly put in and taken out of the gap between the measuring-head electrodes, so that the capacitance between the electrodes is varying at the frequency of insertion of the yarn. This insertion could be done manually at slow speeds or mechanically at high speeds, and, in practice, it is convenient to use the frequency of the mains supply (50 c/s).

(3) VIBRATING-GUIDE INSTRUMENT

Fig. 7 is a block schematic of the instrument which combines the vibrating guide and the self-balancing principle. This is the subject of British and other patent applications.

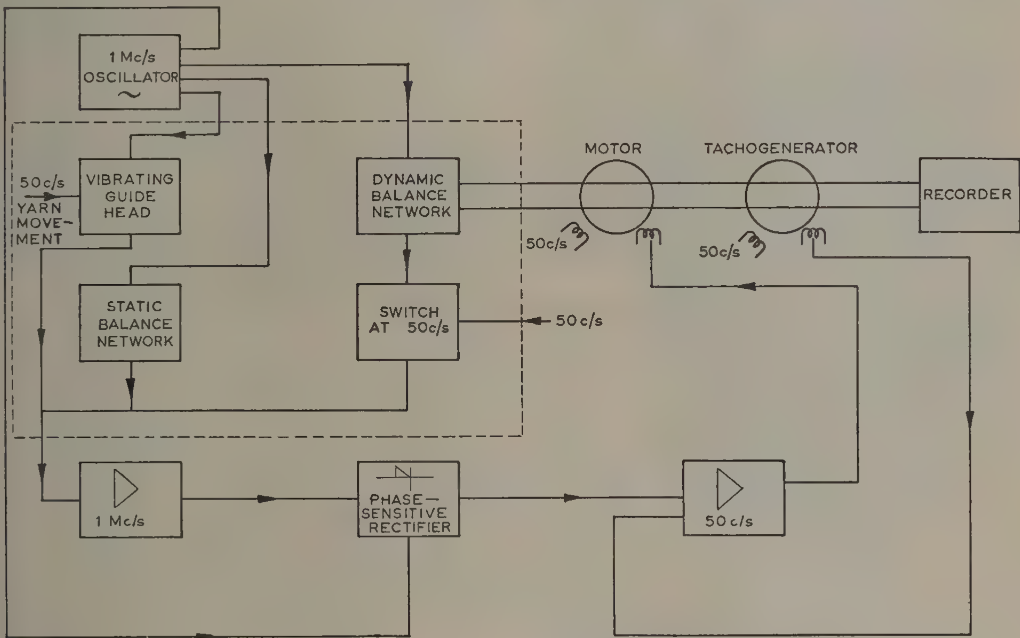


Fig. 7.—Block schematic of the vibrating-guide instrument.

A normal amplifier and oscillator for 1 Mc/s operation has been employed with the bridge circuit. Special bridge transformers using a Ferroxcube core have been constructed to feed the bridge, but it has been found most convenient to use a commercially available wide-band balance-to-unbalance 75-ohm transformer. The bridge receives 5 volts from each half of the secondary winding. The amplifier has an overall gain of over 100 000 and is designed to have as low a phase change as possible over the pass band. A phase-sensitive detector is used, fed with a suitably phased reference signal, so that an output is obtained only with a signal due to capacitive unbalance.

The yarn is guided in and out of the electrodes by two vibrating ceramic pins with V-slots ground in the top, and these guides are driven by a moving-coil vibration generator fed via a suitable transformer from the mains supply. Two stationary ceramic guides are also fitted to feed the yarn running on to the vibrating guides. When the vibrating guides are in their highest position the yarn is clear of the electrode gap, and in their lowest position it is between the electrodes. The capaci-

tance between the two electrodes is therefore varying at the mains supply frequency, and the amplitude of the variation is a measure of the yarn denier.

A static balance network is used to balance the fixed inter-electrode capacitance. A 50 c/s switch is synchronized with the movement of the yarn so that the dynamic balance network is connected into the bridge at the same time as the yarn is inserted between the electrodes. When the capacitance change due to the yarn is not completely balanced by the dynamic balance network, the input to the amplifier will be modulated at 50 c/s, and the output from the phase-sensitive rectifier will contain a 50 c/s component. This is amplified by a 50 c/s servo amplifier and controls an a.c. motor to rebalance the dynamic part of the bridge. The reference winding of this motor is fed with current phased at 90° to the vibrator and switch. The servo motor is also mechanically coupled to a moving-pen carriage which continuously records on a chart the denier of the yarn. The system is stabilized by velocity feedback from a tachometer generator attached to the motor. The feedback voltages are fed into an early stage of the 50 c/s amplifier.

It will be seen that the system is similar in many respects to the self-balancing potentiometer recorder,[14] which has become the standard instrument for recording thermocouple e.m.f.'s, etc., and which is available in many forms. The potentiometer slide wire of such a recorder is replaced by a differential variable capacitor, and the d.c. potentiometer network is replaced by a 1 Mc/s oscillator, bridge circuit, amplifier and phase-sensitive rectifier. This instrument was built around the framework of a standard potentiometer recorder.

this capacitor by eqn. (3). When the dynamic bridge is balance[d] point A in Fig. 6 will be at earth potential and can be sho[rt-] circuited to earth without affecting the static bridge balan[ce]. If the dynamic bridge is unbalanced, short-circuiting A to ear[th] at 50 c/s will produce a 50 c/s modulation, which, in turn, [is] amplified and used to restore balance.

The first short-circuiting switch used was a high-speed rela[y], but later a silicon-diode switch was found more convenient.

It is desirable to have a number of ranges on the instrume[nt] so that measurements may conveniently be made on yarns [of] differing nominal deniers. At the same time operation of t[he] range switch should not alter the static balance of the brid[ge], i.e. the range switch should modify α in eqn. (6) but leave [β] unchanged. When C_x is small, as it is in practice, β is indepe[n]dent of C_x. The range switch S_2 changes α by switching [in] alternative values of C_3. Small adjustments of δC_x for ea[ch] range are made by small adjustments to C_4. This does n[ot] affect β because C_4 is small compared with C_7.

The circuit of Fig. 8 is designed to use 1 in electrodes separat[ed] by 0·02 in. A capacitance change of $0{\cdot}1 \times 10^{-3}$ pF p[er] denier is obtained with this head. The balance control is [a] 50–0–50 pF differential capacitor giving a total change [of] $(C_2 - C_1)$ of 100 pF. The static balance control is a 25–0–25 [pF] differential capacitor. Full movement of this control cor[re]sponds to 1 pF at the head, and β has the constant value of 1/1[0]. This determines the values of C_x, C_3 and C_4, corresponding [to] the five ranges provided on the instrument. Another secti[on] of the range switch S_2 connects alternative values of R_5 [in] parallel with the amplifier input tuned circuit, so that the chan[ge]

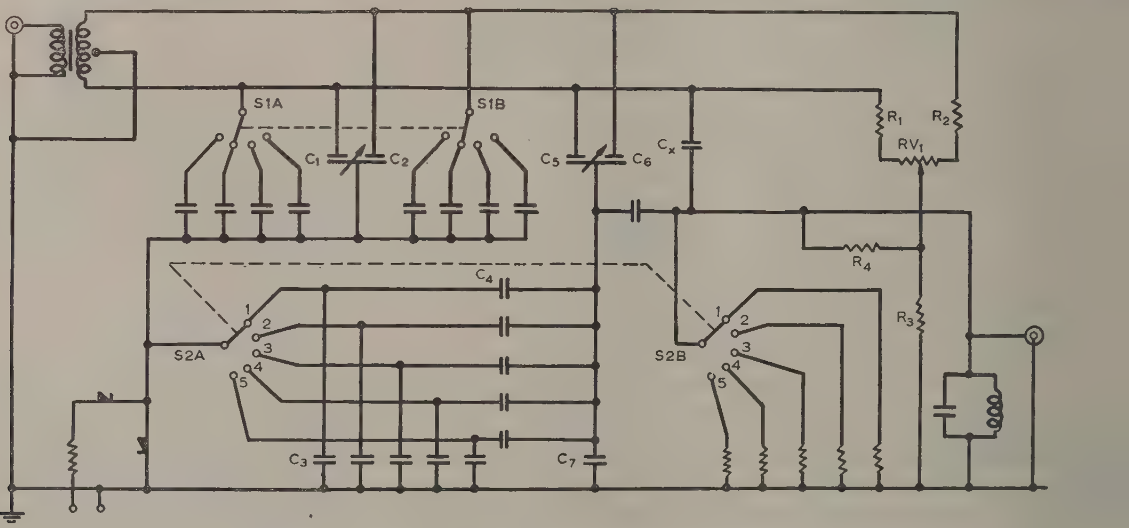


Fig. 8.—Bridge circuit of the vibrating-guide instrument.

The bridge circuit used is shown in Fig. 8, and is basically a combination of two capacitance balance controls acting in series as shown in Fig. 6 with a parallel conductance balance network as shown in Fig. 5. The potentiometer RV_1 and the differential variable capacitor, C_5, C_6, form the static conductance and capacitive balance controls, respectively. Their setting is not critical so long as a constant unbalance signal does not overload the 1 Mc/s amplifier and phase-sensitive detector. They are set for a given head and need not subsequently be adjusted. The dynamic balance control C_1, C_2 is another differential variable capacitor driven by the a.c. servo motor. The capacitance changes at the head are linearly related to the setting of

in input voltage corresponding to a given rotation of the balan[ce] capacitor remains constant. Without these resistors, the ope[n-]loop gain of the complete 50 c/s servo mechanism would chan[ge] with the different denier ranges, and it would not be possible [to] maintain a suitably damped response on all ranges.

It is often convenient to examine small variations in deni[er] on an expanded scale, and S_1 provides a suppressed-zero facili[ty] so that it is possible to use the next more sensitive range. Thr[ee] switch positions are provided on S_1 giving 0%, −50% a[nd] −100% suppressed-zero ranges. Additional capacitors a[re] added in parallel with the two parts of the dynamic balan[ce] capacitor to bring the value of $(C_1 + C_2)$ to 250 pF on all range[s].

o that the value of C_x and the constant calibration are not hanged when the suppressed-zero facility is used.

The 1 Mc/s oscillator, amplifier and phase-sensitive detector re conventional circuits and are made on a common sub-hassis, connected to the bridge, 50 c/s amplifier and power upplies by plugs and sockets. The amplifier uses a low-noise iigh-slope r.f. pentode as the input stage and two r.f. pentodes n the subsequent stages. Gain control is applied to the sup-ressor grids of the valves in preference to amplitude limiting, ince this was found to give less detuning of the circuits. The hase detector uses two germanium diodes, and the 50 c/s omponent of the rectified signal is transformer coupled to the nput stage of the 50 c/s amplifier, which consists of an *RC* oupled pentode. This input stage is followed by two triode mplifying stages feeding a single-ended power-amplifying stage, vhich drives the control winding of the servo motor through a tep-down output transformer. The velocity-feedback signal rom the tachogenerator coupled to the motor is fed into the nput of the second amplifying stage via a potentiometer which s used to adjust the instrument damping.

(4) OPERATIONAL PERFORMANCE

A vibrating-guide denier tester has been in operation for over a year. It has proved to be simple and reliable in operation, nd the calibration has remained quite constant. The static alance of the bridge has proved to be uncritical, and both the apacitance and conductance controls seldom require adjustment.

Long-term zero drift of the instrument is very small, being ess than 0·1 denier on the 100-denier range. No screening has een found necessary on the measuring head, as, unlike the lirect capacitance-measuring type of instrument, the head is naffected by stray capacitances.

The restricted frequency response imposes the limitation that nly long-wavelength irregularities can be observed at high yarn-vind-up speeds. The response speed of the instrument is approxi-nately 1 sec for the full-scale deflection of 10 in. In practice, ariations at 1 c/s of about 1 in peak-to-peak scale amplitudes r 2 c/s of about ¼ in peak-to-peak scale amplitudes can be ollowed when the damping is adjusted for negligible overshoot. The error in the response of the instrument to short-wavelength lenier variations will depend on the amplitude of the variations nd the sensitivity range used in addition to the wavelength of he variations and the yarn wind-up speed.

(5) APPLICATION CONSIDERATIONS

The prototype instrument described has been used for investi-gational purposes in nylon-yarn manufacture, involving a large number of identical production units running in parallel. Because of the reliability and robustness of the instrument urther consideration has been given to its wider application.

For production quality control the main interest is in the nean denier over several thousand metres of yarn, but certain horter-term variations such as those caused by a damaged xtrusion pump also require to be monitored. With extrusion nachines which employ mechanical drives common to a number of extrusion positions it is only possible to effect automatic orrective action based on the average yarn conditions. We nay thus envisage a number of detecting heads—one for each hread line—whose signals are continuously averaged before eing fed to a single measuring unit, which is designed also to orm an error signal for power amplification before this is applied o a motor speed-control unit, thus completing the control loop. Alternatively, if the machine employed individual drives for either he extrusion pumps or the wind-up, the control loop could be pplied to each thread line.

At present, the extrusion characteristics are such that the schemes outlined above, even if practicable, are likely to be of doubtful economic value. More consideration has therefore been given to the use of a detecting head for each thread line, permanently coupled to some central location where either manual or automatic scanning can take place in conjunction with a measuring, alarm and presentation unit.

In view of the requirements for a relatively large number of detecting heads, the production design of this component assumes some importance. The principle features which must be pre-served are those affected by the permanence of the electrical and mechanical properties of the constructional materials. The dimensional stability of certain stainless steels and the electrical characteristics of ceramic insulation can be exploited here, utilizing a symmetrical form of construction allowing for com-plete screening and short connections to the coaxial outlet sockets. Most of these features can be seen in Fig. 9, where

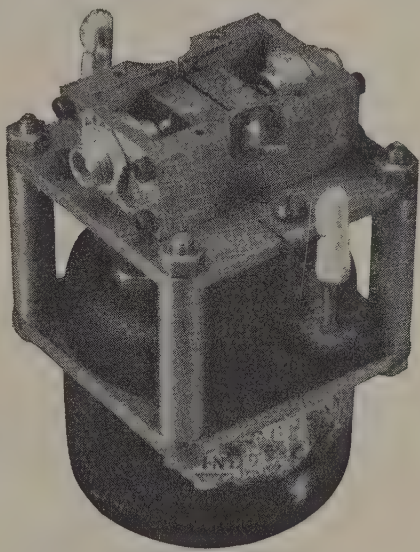


Fig. 9.—The detecting head.

one screening cover plate has been removed to show the internal electrode arrangement. In this design a gap of 0·02 in is used, which gives a capacitance change of approximately 0·1 mpF per denier and permits a useful range of yarns to be accom-modated before the non-linearity between capacitance change and denier becomes excessive. No simple and reliable method has yet been found for trimming a number of such detecting heads to a fixed standing capacitance. In any case it may be more convenient to achieve this, along with cable capacitance balancing, at a central point near the scanning switch.

The vibration generator is a standard component giving an amplitude of $\pm 0{\cdot}05$ in, and its mounting together with the bridge-piece holding the grooved guides has to be sufficiently rigid to preserve the relationship required between the yarn path and the electrode gap. The fixed guides should be adjust-able to obtain the minimum of displacement from the straight thread line, this, in turn, minimizing thread-line tension variation.

The location of the detecting head on the extrusion machine presents some problems, the final choice being a compromise involving such factors as:

(*a*) Accessibility for introducing the thread line and for main-tenance purposes.

(*b*) Moisture and temperature equilibrium of the thread line at the point of measurement.

(*c*) Effect of tension variations due to the imposed 50 c/s transverse motion of the thread line.

(*d*) Protection against excessive accumulation of dust and other foreign matter in the capacitance gap, which would affect the apparent denier.

(*e*) Retention of sufficient clearance to allow for the normal thread-line manipulations.

These characteristic difficulties are reduced in cases where the machine is initially designed to accommodate such equipment, and this emphasizes one of the factors which should influence machine design and development.

The vibration-generator power supply may obviously be a single run of twin insulated wire looped into the terminal boxes on each of the heads and carried in conduit; or mineral-insulated metal-sheathed cable may be used cleated directly to suitable machine members. The pairs of coaxial cables from the capacitance electrodes should be run separately in cable trunking, well protected from mechanical damage and terminated in a balancing connection box in or near the scanning switch.

The scanning switch itself may be a conventional manually operated or automatic stepping rotary switch with low loss and leakage characteristics. Its rotational speed, if automatic scanning is employed, need only be slow, the measuring time spent for each detecting head being related to the process characteristics and the number of thread lines served by the central measuring unit. A time of the order of 20 sec per point would appear to accommodate normal process-control requirements, if provision is also made for overriding the scanner at any time to obtain a measurement of one particular thread line on demand.

The presentation on the prototype is obtained by a servo-operated recorder with a 10 in-wide chart, and this lends itself directly to any other form of presentation by means of a rotary digitizer attachment.

The instrument has proved useful as an investigational tool having adequate accuracy, stability and frequency response and being capable of extended use as a means of routine quality control either by open-loop monitoring or closed-loop automatic control.

(6) ACKNOWLEDGMENTS

The authors are indebted to British Nylon Spinners, Ltd., for permission to publish the paper, and to several of their colleagues for assistance. In particular, they acknowledge practical work by Messrs. R. E. Ricketts and G. A. Stickler.

(7) REFERENCES

(1) MACK, C.: 'The Field due to an Infinite Dielectric Cylinder between Two Parallel Conducting Plates', *British Journal of Applied Physics*, 1955, **6**, p. 59.

(2) MACK, C.: 'The Capacitance of a Parallel Plate Condenser with an Anisotropic Dielectric Cylinder in Torsion between its Plates', *Philosophical Magazine*, 1951, **4**
p. 428.

(3) MACK, C.: 'The Capacitance of a Parallel Plate Condense with an Anisotropic Dielectric Cylinder in Torsio between its Plates', *Shirley Institute Memoirs*, 1951, **2**
p. 191.

(4) MACK, C.: 'Some Factors affecting the Change in Capacit of a Parallel Plate Condenser due to the Insertion o Yarn', *Journal of Textile Institute (Transactions)*, 195 **46**, p. 500.

(5) BOYD, G. N.: 'An Electronic Instrument for Measurin Weight Variations in Slivers, Rovings and Yarns', *ibid* 1949, **40**, p. 407.

(6) WALKER, P. H.: 'The Electronic Measurement of Slive Roving and Yarn Irregularity with special reference t the Use of the Fielden Bridge Circuit', *ibid.*, 1950, **4** No. 7, 446.

(7) FIELDEN, J. E.: British Patent No. 619534, 1949.

(8) ZELLWEGER A.G.: Apparate und Maschinenfabriken Ust British Patent No. 638365, 1950.

(9) SHIRLEY INSTITUTE: 'The Effect of Moisture upon th Accuracy of Capacity Type Regularity Testers', *Shirle Institute Bulletin*, 1956, **29**, p. 57.

(10) HEARLE, J. W. S.: 'Capacity, Dielectric Constant and Powe Factor of Fiber Assemblies', *Textile Research Journa* 1954, **24**, p. 307.

(11) WINDLE, J. S., and SHAW, T. T. M.: 'Microwave Dielectr Measurements on Single Fibres', *ibid.*, 1955, **25**, p. 865.

(12) STANDARD TELEPHONES AND CABLES, and BLUMLEIN, A. D British Patent No. 323037. 1928.

(13) CLARK, H. A. M., and VANDERLYN, P. B.: 'Double-Rati A.C. Bridges with Inductively-Coupled Ratio Arms *Proceedings I.E.E.*, Paper No. 742, January, 1949 (**9** Part III, p. 189).

(14) WATTON, W. L., and PEMBERTON, M. E.: 'A Direc Capacitance Aircraft Altimeter', *ibid.*, Paper No. 74 January, 1949 (**96**, Part III, p. 203).

(15) WOJCIECHOWSKI, B. M.: 'Capacitance Gauge Checks Cab Sheath Thickness', *Electronics*, 1954, **27**, p. 134.

(16) CALVERT, R.: 'Industrial Applications of Transform Bridge Circuit', *British Communications and Electronic* 1956, **3**, p. 180.

(17) KIRKE, H. L.: 'Radio Section: Chairman's Address *Proceedings I.E.E.*, 1945, **92**, Part III, p. 2.

(18) YOUNG, C. H.: *Bell Laboratories Record*, 1946, **24**, p. 44

(19) YOUNG, C. H.: 'Measuring Inter-Electrode Capacitance *Tele-Tech*, 1947, **6**, p. 68.

(20) YOUNG, C. H.: 'Measuring Inter-Electrode Capacitances *Technique*, 1947, **22**, p. 343.

(21) ELECTRONIC TUBES, LTD.: 'R.F. Capacitance Bridge Internal Report No. LR 140.

21.314.63 : 621.382.333.3

The Institution of Electrical Engineers
Paper No. 3281 E
Sept. 1960

THE DESIGN OF CONTROLLED RECTIFIERS USING TRIODE TRANSISTORS

By E. E. WARD, Ph.D., Associate Member.

(*The paper was first received* 24*th October*, 1959, *and in revised form* 11*th February*, 1960.)

SUMMARY

The paper deals with polyphase rectifiers in which the switching lement is a complementary combination of a *p–n–p* and an *n–p–n* ransistor. The stability conditions for the open and closed states of his switch are investigated graphically, leading to the calculation of the oltage drop when conducting. There is a sharp upper limit to the ermissible current. Tests are described on a 3-phase controlled ectifier.

LIST OF SYMBOLS

I_1 = Emitter current of transistor 1.
I_{b1} = Base current of transistor 1.
I_4 = Collector current of transistor 1.
I_{c2} = Collector current of transistor 2.
I_3 = Base current of transistor 2.
I''_{c01} = Saturation current of transistor 1, common collector.
I'_{c02} = Saturation current of transistor 2, common emitter.
I_5 = Fictitious current flowing in R_3.
I_6 = Fictitious current injected at point D.
$I_0 = I''_{c01} + \alpha''_1 I'_{c02}$.
I_{01} = Saturation current of junction R_1.
k = Boltzmann's constant.
e = Electron charge.
R_1 = Resistance in collector circuit of transistor 1.
R_2 = Resistance in base circuit of transistor 2.
$R_3 = R_1$; it closes the base circuit of transistor 2 when loop is opened.
R_4 = Resistance of junction R_1.
T = Absolute temperature.
T_1 = Temperature of *p–n* junction forming R_1.
T_2 = Temperature of emitter junction of transistor 2.
V_{cb1}, V_{cb2} = Collector–base voltages.
V_{ce1}, V_{ce2} = Collector–emitter voltages.
V_{be1}, V_{be2} = Base–emitter voltages.
α_1, α_2 = Current amplification factors, common base.
α'_1, α'_2 = Current amplification factors, common emitter.
α''_1, α''_2 = Current amplification factors, common collector.

(1) INTRODUCTION

The paper deals with polyphase rectifiers wherein the output voltage is controlled by the instant of firing the switching elements. Grid-controlled arc rectifiers, ignitrons, thyratrons and similar devices are widely used in these circuits and a controlled semiconductor diode of *p–n–p–n* construction is now becoming available which will offer lower voltage drop, less weight and smaller size than these switches at present in use and needs no waiting period to warm up before taking up its load. Fig. 1 shows a circuit which is typical of rectifiers of this kind.

The present paper investigates quantitatively the circuit proposed by Shockley[1] and applied by Deuitsch and Paz[2] in which two triode transistors operate as a controlled switch. This arrangement behaves like the *p–n–p–n* diode in some respects, and under present commercial conditions the cheapness of the components offsets the fact that four are needed. Since, moreover, transistors of types which are already widely available can be used, the circuit is an attractive alternative to the *p–n–p–n* controlled diode.

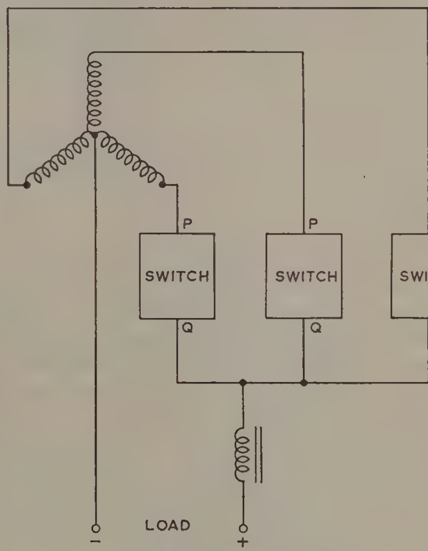


Fig. 1.—Rectifier circuit.

The triode junction transistor can control considerable power when used as a switch since its leakage current when open and its voltage drop when closed are each small. However, the directions of its currents are such that no passive circuit-element will cause the transistor to lock itself in the fully open or fully closed state, and an auxiliary transistor is used to give large amplification in a regenerative loop so that these two stable states are obtained. The circuit which has been studied is shown in Fig. 2; it uses transistors of opposite polarities and will be known as a 'complementary switch'.

(2) COMPLEMENTARY SWITCH

In the circuit of Fig. 2 the switching transistor 1 is taken for convenience as having *p–n–p* polarity; the main load current flows from its emitter to its collector. Transistor 2 controls the base current of transistor 1 and may therefore be of lower current rating; it must be of *n–p–n* polarity. The analysis which follows will be based on these polarities; it may be applied to a circuit in which the transistor polarities are interchanged by reversing the directions of all potentials and currents.

In a controlled rectifier such as that shown in Fig. 1, the switch elements must be capable of three stable states, as follows:

A state; reverse voltage; switch open.
B state; forward voltage before triggering; switch open.
C state; forward voltage during conduction; switch closed.

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.
Dr. Ward is in the Electrical Engineering Dept., University of Birmingham.

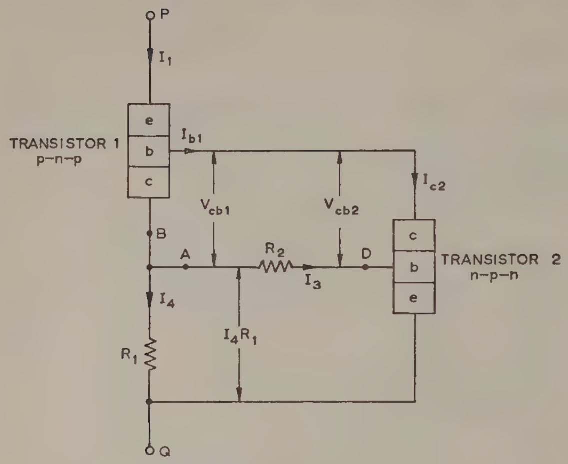


Fig. 2.—Circuit of switch.

In the B and C states the switching transistor 1 is polarized in its normal working direction; it is, however, equally capable of conducting, though with reduced performance, under the reversed potentials of the A state, the emitter and collector interchanging their functions, and in the A state the base must be so controlled as to prevent this conduction. The three states are established by the control transistor 2.

A State.—Terminal P negative. The emitter of transistor 1 acts as a collector, and the collector as an emitter; to prevent current flow, the base must be at collector potential. The reverse voltage is sustained by the emitter junction of transistor 1 and the other electrodes of both transistors are at the potential of terminal Q, as shown in Fig. 3. The stressed emitter junction of transistor 1 passes its small reverse saturation current and the breakdown of this junction sets the limit of reverse voltage which the switch will withstand in the A state.

B State.—Terminal P positive. The electrodes of th transistors behave normally as they are labelled in the Figur To prevent conduction, the base of transistor 1 must be a the potential of terminal P, and that of transistor 2 at th potential of Q. Thus both collector junctions are polarize in the normal working sense, i.e. the reverse sense, and bot must sustain the applied voltage of the B state. Both junctior pass their saturation currents, but these currents are stab if R_1 is below a certain critical value which depends on th temperatures of the junctions. The electrode potentials ar shown in Fig. 3.

C State.—Terminal P positive. When the rectifier is work ing a trigger pulse is applied to the switch in the B state an produces a transition to the C state, as will be describe later; both transistors then act as closed switches. To kee the voltage drops between collectors and emitters low, bot transistors must carry heavy base currents and their collecto potentials must lie between those of their respective bases an emitters. Fig. 3 shows how this is possible. Since th transistors are of opposite polarities and are connected i opposite senses in the circuit, the voltage drop I_4R_1 produce by the load current serves to excite both transistors; it feeds heavy base current to transistor 2, carrying its base positiv and this transistor, acting as a closed switch, brings the bas of transistor 1 near to the potential of terminal Q; at th same time the collector of transistor 1 is made positive wit respect to terminal Q by the potential I_4R_1.

This state is also stable provided that R_1 exceeds a certa critical value which depends on the values of α_1 and α and on the minimum load current at which the switch required to remain closed. The potentials marked in Fig. were measured for a load current of just over 2 amp. Th Figures show that transistor 1 accounts for 760 mV, 430 m in the conducting path between emitter and collector and

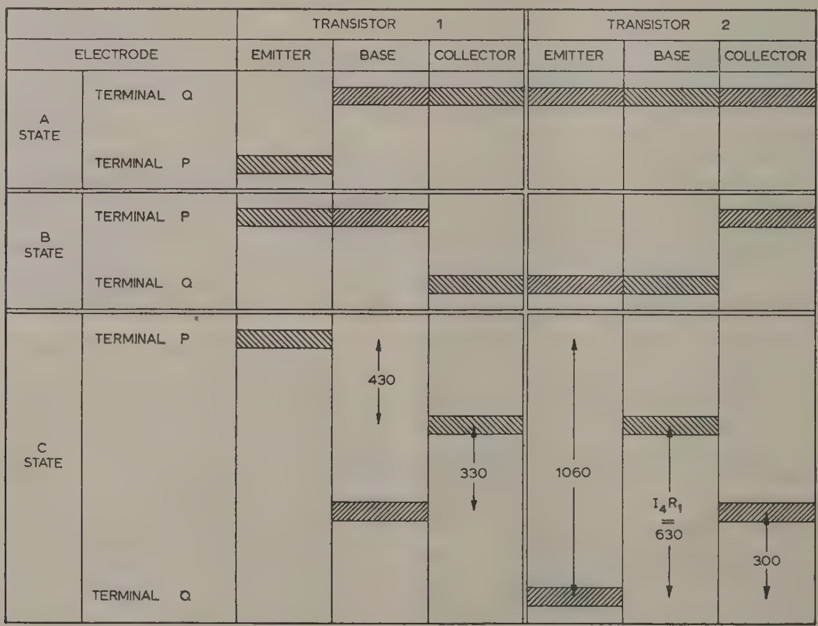


Fig. 3.—Relative electrode potentials in millivolts.
Upward direction positive.

further 330 mV as the price of this reduction in the emitter–collector potential. The heavy base current of transistor 1 when passing through transistor 2 causes an emitter–collector potential of 300 mV, and this brings the total to 1·06 volts.

(3) TRIGGERING OF THE SWITCH

The circuit of Fig. 2 may be stable in both the open and closed states and may be triggered from one to the other.

(3.1) Stability of the B State

Assuming R_2 to be zero and breaking the closed loop at A, an open control chain results as shown in Fig. 4, where a resis-

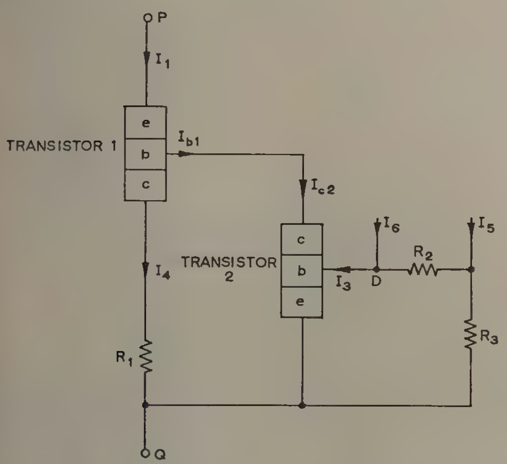


Fig. 4.—Open-loop circuit of switch.

tance R_3 has been added equal in value to R_1 and nodal currents I_5 and I_6 flow as shown. The potential between terminals P and Q will be assumed to be unaffected by the small currents which flow, the collector of transistor 2 and the base of transistor 1 being at the potential of P, and I_6 will be taken to be zero.

Then $$I_4 = \frac{1}{1-\alpha_1}\left[\left(\frac{\alpha_2}{1-\alpha_2}\right) \times I_3 + I'_{c02}\right] + I''_{c01}$$

i.e. $$I_4 = \alpha''_1[\alpha'_2 I_3 + I'_{c02}] + I''_{c01} \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

This makes I_4 a linear function of I_3; in practical circuits α''_1 and α'_2 fall as I_4 rises and $\alpha''_1 I'_{c02}$ is much smaller than I'_{c01}; the relationship is shown in Fig. 5 by the dotted line. The upper asymptote of I_4 is that determined by the external circuit, and is independent of I_3. The intercept on the I_4 axis has the value $(\alpha''_1 I'_{c02} + I''_{c01})$ when I_3 is zero.

However I_3 is close to being an exponential function of the voltage drop I_5R_3, or if R_3 is given

$$I_3 = R_3 \mathrm{f}(I_5) \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (2)$$

This relation is represented by the solid line in Fig. 5. When the loop is closed $I_4 = I_5$ and operation must be at one of the points of intersection at which eqns. (1) and (2) are both true. Of the three intersections, P_3 is unstable, and P_1 and P_2 are stable; P_1 represents the B state and P_2 the C state. At P_1 the current I_4 has the value OB, exceeding the intercept OA by the amount $\alpha''_1\alpha'_2 I_3$, where BP_1 represents I_3. Eqn. (2) shows that the abscissae I_3 are directly proportional to R_3, which is equal to R_1

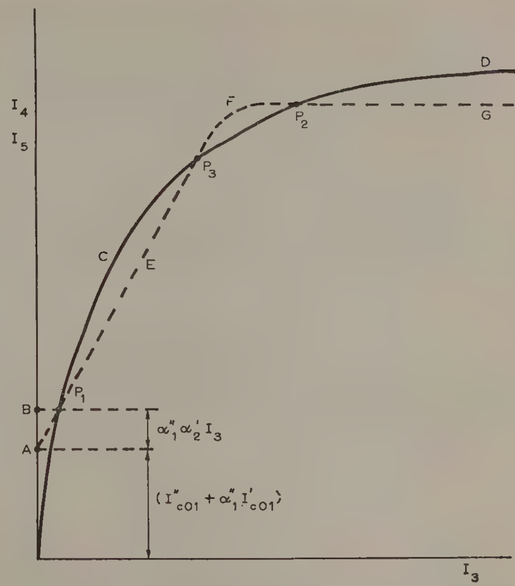


Fig. 5.—Currents flowing in circuit of Fig. 4.

---- Eqn. (1). —— Eqn. (2).

and which, as it increases, moves the curve OCD away from the I_4 axis. In this way P_1 and P_3 close up; in the limiting condition when they coincide, P_1 becomes unstable and operation can be only at P_2 in the C state. Thus the B state cannot be established if R_3 (being equal to R_1) equals or exceeds the value which makes the curve AEFG an external tangent to OCD. The diagram shows that the critical value of R_1 is reduced by increasing the slope of the line AEF, i.e. the product $\alpha''_1\alpha'_2$, and also by increasing the intercept OA representing $[\alpha''_1 I'_{c02} + I''_{c01}]$, which varies markedly with temperature. Fig. 6 shows a family of

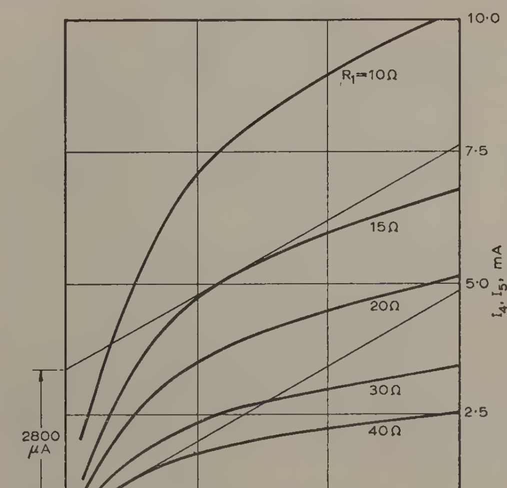


Fig. 6.—Stability condition for B state.

Type V10/15A transistor with experimental *n–p–n* sample.

such curves measured for a *p–n–p* transistor V10/15A and an experimental *n–p–n* sample over a range of values of R_1. The intercept OA represents 550 μA and the tangential curve is for $R_1 = 40$ ohms, which is in agreement with experiment.

(3.2) Triggering Current

For operation as a controlled switch R_1 is less than the critical value corresponding to the intercept OA, and the B state is thus stable. With R_2 still zero the circuit may be triggered to the C state by injecting an external current at the point B in Fig. 2 in the same sense as I_4, thus, in effect, increasing the saturation current

$$I_0 = (\alpha_1'' I_{c02}' + I_{c01}')$$

and moving the curve AEF parallel to the I_4 axis until it becomes tangential to the curve OCD. The current needed is exactly that represented by the shift along the I_4 axis and it increases as R_1 is reduced. Fig. 6 shows that when $I_0 = 550\,\mu$A a shift of 2·80 mA is needed to make the control curve tangential to the curve $R_1 = 15$ ohms, and the measured value is in agreement.

The necessary triggering current may be reduced by an order of magnitude if R_2 is made finite. For currents typical of the B state, the input d.c. resistance of transistor 2 may be 1 000 times greater than R_1. If R_2 has a value close to the geometric mean of the transistor resistance and R_1, its presence will cause only a slight fall in I_3. If, now, the external triggering current I_6 be applied at point D of Fig. 2, the corresponding point being marked D in Fig. 4, little change will occur in the current carried by R_1 but I_3 will be due to a potential

$$[I_6(R_2 + R_3) + I_5 R_3] = R_3\left[I_5 + I_6\left(\frac{R_2 + R_3}{R_3}\right)\right]$$

and the I_4 ordinates of the curve OCD are therefore reduced by $I_6(1 + R_2/R_3)$. Thus since $R_3 = R_1$ a given triggering current injected at D will produce a relative shift of the curves $(1 + R_2/R_3)$ times greater than if it were injected at B. These reduced values of triggering current are again in agreement with experiment.

(3.3) Stability of the C State

The C state is represented in Fig. 5 by point P_2, the value of I_4 on the asymptote FG being determined by the external circuit; $(I_4 + I_{b1})$ is, in fact, the load current carried by the closed switch. Fig. 5 shows that the positions of P_1 and P_3 are determined by the positions of the intersecting curves, and these, in turn, by R_1, I_6 and the product $\alpha_1''\alpha_2'$; of these α_1'', and α_2' vary slightly with the load current whilst R_1 and I_0 are independent of it. Thus a decrease in load current brings the asymptote FG nearer to the I_3 axis, causing P_2 to approach the fixed point P_3. In the limit they coincide, the C state becomes unstable and the switch moves to P_1, the B state. For a given position of the curve OCD representing a given value of R_1, there is therefore a minimum load current below which the switch transfers to the B state, and this limiting current is the value which makes the curve EFG an internal tangent to OCD. An example is shown in Fig. 7 plotted for the same transistors as Fig. 6; a load current of 16 mA is just sufficient to keep the switch closed when $R_1 = 10$ ohms.

(3.4) Stability of the A State

Since the inverse voltage polarizes the emitter junction of the switching transistor in the reverse direction and leaves all other electrodes at the potential of terminal Q the switch is proof against operation by triggering currents of either polarity whether applied at point B or D in Fig. 2.

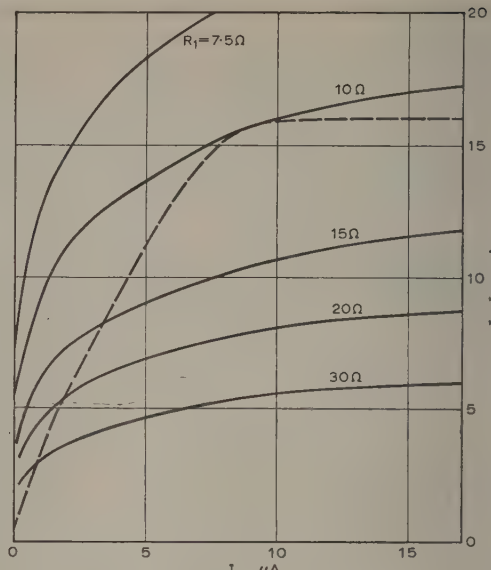


Fig. 7.—Stability condition for C state.
Transistors as Fig. 6.

(4) CONDUCTING CONDITION

(4.1) The Resistor R_1

The potential between terminals P and Q when the switch carrying its rated current must be known in order to allow th efficiency of the complete rectifier to be predicted; further th currents in and potential drops across each of the two transisto must be kept within their safe working limits; thus an analys in detail of conditions in the C state is essential. Difficulti are caused by the large ratio of the currents in the C state those in the B state, and no system of co-ordinates has be found which overcomes them all; however, it is convenient combine charts such as Figs. 6 and 7 by plotting I_4 in line ordinates and I_3 in logarithmic abscissae. These co-ordinat have the further advantage that eqn. (2) would be represented a straight line if it were exactly exponential. A hypothetic chart of this kind is shown in Fig. 8, where the curves show th

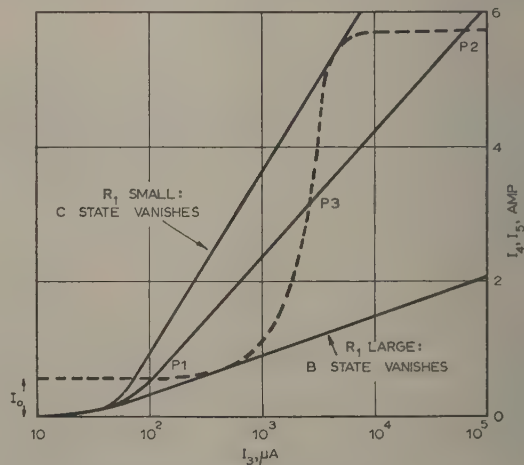


Fig. 8.—Bounds of stability.
---- Eqn. (1). —— Eqn. (2).

ideal behaviour which would be found if the p–n junctions were free of ohmic resistance and where the saturation current I_0 has been shown to be disproportionately large in order to show the trends of the curves at low currents.

Fig. 9 is a chart in the same co-ordinates showing measured

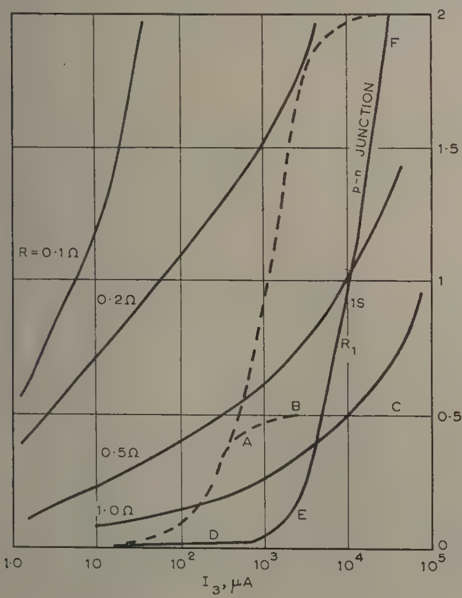


Fig. 9.—Stability condition for C state.
Types V30/20P and OC139 transistors.

values of types V30/20P and OC139 transistors; the dotted curve, as in Fig. 5, represents eqn. (1); similarly, the solid curves represent eqn. (2); the methods described for predicting stability and triggering are all applicable; R_2 is again zero. The curves show clearly that conflicting requirements arise in choosing R_1; a small value will fail to hold the C state at low currents; for example, the line ABC representing a low load current of $I_4 = 0{\cdot}5$ amp has no C state intersection with the line $R_1 = 0{\cdot}5$ ohm; but if R_1 were increased to 1·0 ohm which gives a definite P_2 point for a minimum load of 0·5 amp, the value of I_3 at the full load 2·0 amp would be destructively high; the lines do not, in fact, intersect within the bounds of Fig. 9. To carry such currents safely the auxiliary transistor would be larger than is needed for controlling the switch.

An alternative to this costly choice is to hold I_3 roughly proportional to I_4. This may be done by using as the resistor R_1 a p–n junction[4] polarized in the forward direction, preferably shunted by an ohmic resistor to limit the rise of resistance at low currents. Since this junction diode is not subject to reverse voltage a unit of low commercial quality will suffice; and with an auxiliary transistor of the reduced rating which it allows, it should prove cheaper than the alternative p–n–p transistor of much greater rating. Such a modification makes R_1 a function of I_4 and their product results in a value of I_3 shown by the curve DEF of Fig. 9. It will be seen that I_3 and I_4 are roughly proportional, and for ideal junctions of which the p–n junction is at a temperature T_1 and transistor 2 at T_2, I_3 will be proportional to $I_4{}^{T_1/T_2}$. This exponent is not likely to differ greatly from unity; for example, if $T_1 = 15°$C and $T_2 = 75°$C, on the absolute scale $T_1/T_2 = 0{\cdot}83$.

In low-power devices where high efficiency is not important the same result may be obtained by using a fixed ohmic resistor for R_1 and a finite value of R_2; this value should equal the emitter–base d.c. resistance of transistor 2 at about 50% of full load. Above this load I_3 will approach the value I_4R_1/R_2 as required; the voltage drop I_4R_1 is, however, in the main recifying circuit and is larger than the potential required by transistor 2, so that the consequent losses would be acceptable in low-power equipment only.

(4.2) Voltage Drop when Conducting

Fig. 2 shows that transistor 1 is in the common-collector and transistor 2 is in the common-emitter connection. I_3 is negligible in comparison with I_4, which is the most convenient independent variable. The load resistance in the collector circuit of transistor 2 is the base–collector resistance of transistor 1, and the driving e.m.f. in this circuit is the voltage drop I_4R_1. R_1 will be assumed to consist of a p–n junction and an ohmic resistor connected in parallel. Using subscripts 1 and 2 for the two transistors and again assuming that $R_2 = 0$, we have

$$I_4R_1 = V_{cb1} + V_{ce2} \quad . \quad . \quad . \quad . \quad . \quad (3)$$

$$I_{b1} = I_{c2} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (4)$$

$$V_{be2} = I_4R_1 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (5)$$

Eqn. (3) becomes clear on considering the potential diagram (Fig. 3). These equations are to some extent amenable to analysis, and the expressions developed by Ebers and Moll[3] may be substituted into eqn. (3). However, the expressions for V_{cb1} are in poor agreement with the measured values, an example of which is given in Fig. 10, and particularly in the region of

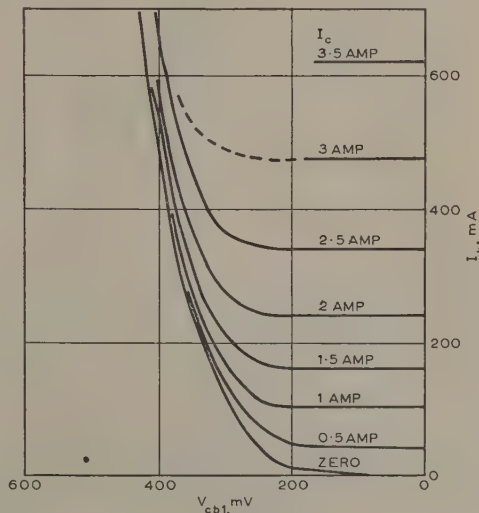


Fig. 10.—Relationship of base voltage and current with common-collector connection.
Type V30/10LP transistor.

operation. The discrepancies seem to be due to the variation of the effective base resistance as a function of the electrode currents and are large enough to exclude an algebraic solution for the operating point. The equations may, however, be solved graphically using the following data:

Graph 1.—Voltage/current curve for junction diode R_1.
Graph 2.—Common-collector curves, emitter current/emitter voltage curves for transistor 1 for various constant base currents.
Graph 3.—Common-collector curves for transistor 1 giving base voltage as a function of base current for various constant values

of collector current (the same curves in terms of emitter current are less convenient). A set of such curves for the V30/10LP transistor is given in Fig. 10.

Graph 4.—Common-emitter curves for transistor 2 giving collector current as a function of collector voltage for constant values of base–emitter voltage.

Taking I_4 as the independent variable, the calculation of the overall voltage drop takes the following stages:

(*a*) For a given value of I_4 graph 1 gives the voltage I_4R_1 across the junction diode R_1.

(*b*) For transistor 2, I_4R_1 is both the collector-circuit supply voltage and also the base–emitter voltage. Hence, choosing the curve for this known base voltage, graph 4 gives the relationship between I_{c2} and V_{ce2}.

(*c*) For a given value of I_4 the behaviour of the load impedance in the collector circuit of transistor 2 may be found in graph 3. Thus by superimposing graphs 3 and 4, so that their voltage axes coincide and their current axes are separated by a voltage I_4R_1, the intersection of the curves for current I_4 gives the operating point. Base potential and base current for transistor 1 are given at this point, as shown in Fig. 11.

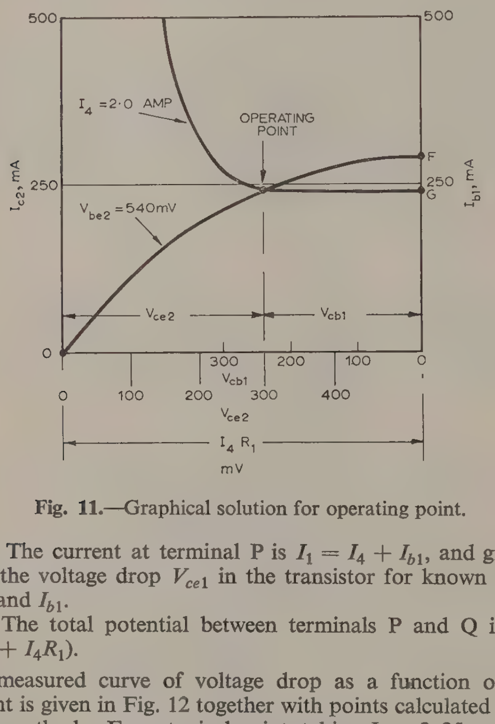


Fig. 11.—Graphical solution for operating point.

(*d*) The current at terminal P is $I_1 = I_4 + I_{b1}$, and graph 2 gives the voltage drop V_{ce1} in the transistor for known values of I_1 and I_{b1}.

(*e*) The total potential between terminals P and Q is then $(V_{ce1} + I_4R_1)$.

A measured curve of voltage drop as a function of load current is given in Fig. 12 together with points calculated by the above method. For a typical point, taking $I_1 = 2{\cdot}25$ amp, and $V_{PQ} = 0{\cdot}94$ volt, the following values may be read from the graphs as explained above:

I_1	2·25 amp
I_4	2·0 amp
I_4R_1	540 mV
V_{ce2}	300 mV
V_{cb1}	240 mV
I_{b1}	250 mA
V_{ce1}	400 mV
$V_{pq} = V_{ce1} + I_4R_1$	940 mV
Dissipation	
Transistor 1	900 mW
Transistor 2	90 mW

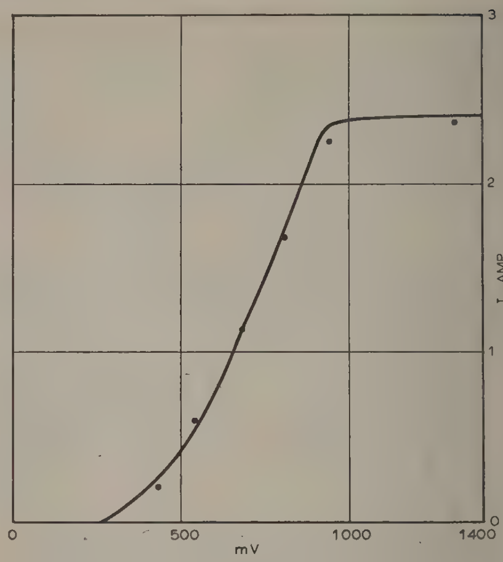


Fig. 12.—Measured and calculated voltage drop for C state.

(4.3) Upper Limit of Current

In order to hold transistor 1 in the closed state so that th voltage drop from emitter to collector will be low, its base mus be held negative with respect to its collector; the constructio shown in Fig. 11 must therefore result in an operating poin lying to the left of the vertical I_{b1} axis and OF must exceed OG Along the line OGF, V_{cb1} is zero and therefore $I_4 = \alpha_1'' I_{b1}$ Thus $OG = I_4/\alpha_1''$.

Turning next to transistor 2, its base is held at a potentia I_4R_1 positive to its emitter whereas its collector potential V_{ce} is less than I_4R_1 over the whole of the diagram except th vertical I_{b1} axis. This transistor also is therefore in the close condition. However, base and collector have the same potentia along OFG and therefore $OF = \alpha_2' I_3$. If R_1 is a *p–n* junctio we may assume $I_3 = KI_4$, giving $OF = K\alpha_2' I_4$. Thus if OF i to exceed OG the product $K\alpha_1''\alpha_2'$ must exceed unity.

The ratios α_1'' and α_2' both fall as I_4 rises, the points F an G close up, and, as a result, the operating point moves toward the I_{b1} axis. In this region the two curves shown intersect at large angle, and there is therefore a sudden rise in the bas potential of transistor 1 leading to a similar rise in the voltag drop across it. This is shown by the measured curve in Fig. 1 and sets an upper limit to the load current of the switch.

(5) TRIGGERING CIRCUITS

Conventional circuits are available which will supply a shor triggering pulse to determine the instant of closing each comple mentary switch, and details may be found in Reference 5. Th base circuit of transistor 2 should be of low ohmic resistanc and the triggering pulse may conveniently be supplied by a low impedance transformer winding connected in the place of R_2, a shown in Fig. 2. No investigations have been made of th possibility of triggering by sinusoidal voltages of variable phas or by direct voltages; both methods are well known (e.g. se Reference 2) and will have their peculiar fields of applications.

(6) INCREASE OF POWER OUTPUT

(6.1) Current

No experiments have been made with transistors connected i parallel; it is common knowledge that current output may b

ultiplied in this way provided that small resistances are connected in each collector lead with the object of equal sharing of the total load.

(6.2) Voltage

The rectifier output voltage at zero firing delay depends only n the ability of the transistors to withstand the voltages applied to them. The conditions of stress are as follows:

(6.2.1.) A State.

No voltages are applied to transistor 2. The peak potential across the emitter junction of transistor 1 is nearly twice the transformer peak phase voltage for single-phase or polyphase star connection.

(6.2.2) B State.

Fig. 3 shows that the voltage between the switch terminals appears across the collector junctions of both transistors; this voltage will approach the transformer peak phase voltage.

In both the A and B states no load currents are flowing; the output voltage of the rectifier may therefore be multiplied by connecting transistors in series, provided that they are stressed equally by the addition of the familiar shunt resistors. However, the complication of providing the control for determining definite B and C states for an assembly of transistors in series suggests that it would be wiser to take several complementary switches complete and connect them in series. This has been done to the extent of three in series; voltage stresses in the A and B states have been equalized by shunt resistors, a separate triggering winding being provided on the pulse transformer for each switch in the chain.

(7) DESIGN POLICY

The following points need special attention:

The loading of the transistors is limited by their mean heat dissipation and also by the mean and peak currents. In a polyphase rectifier each switch is conducting for a fraction only of each cycle, and this increases the peak current which may be allowed for a given mean current and dissipation. A further limit is set by the circuit behaviour as described in Section 4.3.

The cost of the components and the electrical performance are affected by the choice of the *p–n* junction included in R_1. A high voltage drop across this junction will hold the two transistors firmly in their closed states so that V_{ce1} will be small, but the total fall in potential from P to Q may be too large to be acceptable.

The stability of the B state may be controlled using the construction of Fig. 6. The value of R_1 to be used is that given at low currents by the *p–n* junction and the ohmic resistor connected in parallel with it; in practical circuits the conductance of the *p–n* junction may be negligible. The values of I'_{c01} and I'_{co2} should be taken for the highest permitted junction temperatures.

Questions of rating and protection from overload were not included in the investigation, but they are of primary importance since the component transistors of a complementary switch have small thermal time-constants. Solutions of these problems based on experience with diode rectifiers have been published[6, 7] and should be studied.

(8) TEST RESULTS

Complementary switches have been used in a 3-phase rectifier connected according to Fig. 1. The transistors were types V60/30P and OC139; performance data measured on similar components are given in Figs. 9, 10 and 11. Two-ohm resistors were shunted across the *p–n* junctions to form R_1 and the resistors R_2 were replaced by the pulse-transformer output windings. The pulse was roughly triangular, of peak value 0·5 volt and base duration about 1·5 millisec; a rotary phase-shifter was used to give an adjustable firing delay of up to 120°. The star-connected transformer was a standard laboratory unit consisting of three variable transformers ganged on a common shaft, and was supplied at 50 c/s at a voltage well below its rating in order to preserve a sinusoidal waveform. The output direct voltage of this circuit was limited by the voltage rating of available commercial transistors, particularly those of *n–p–n* polarity, to about 11 volts; curve (*c*) of Fig. 13 represents the performance

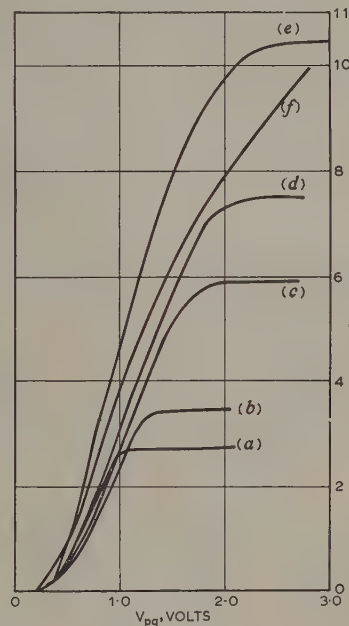


Fig. 13.—Measured voltage drop in C state.

of the switches, and the average direct current was held at about 4 amp. Voltage control by delaying the firing pulse operated in the normal way.

Fig. 13 shows the voltage-drop curves of several switch arrangements:

Curve (*a*).—The switching transistor was of type V60/30P and the controlling transistor was a sample of type OC139 chosen as having a low value of α'. R_1 was a 2-ohm wire resistor in shunt with the collector junction of a V30/10LP transistor. For the collector junction the terminal voltage was in approximate agreement with the equation

$$V = \frac{kT}{e}\log_\varepsilon\left(1 + \frac{I}{I_{01}}\right) + IR_4$$

in which $I_{01} \simeq 170\,\mu\text{A}$

$R_4 \simeq 0\cdot03$ ohm.

Curve (*b*).—The same transistors were used as for curve (*a*); the *p–n* junction was now the emitter junction of the same transistor, giving

$I_{01} \simeq 115\,\mu\text{A}$

$R_4 \simeq 0\cdot08$ ohm

At a current I_4 of 3 amp the effect is to increase the potential fall across R_1 by 210 mV. The Figure shows that the overall voltage drop and the upper limit of current have both risen.

Curve (*c*).—The switching transistor was the same sample of type

V60/30P, the *p–n* junction was the emitter junction of V30/10LP type and the low α' sample of type OC139 was replaced as the control transistor by one of high α'. The voltage drop at low current is still greater than for curve (*a*), but the upper limit of current shows a further considerable rise.

Curve (*d*).—Two type OC139 transistors, each of high α', were connected in parallel. Otherwise the circuit was as for curve (*c*). The voltage drop is lower throughout, and the upper current limit is higher than in any of the previous curves.

Curve (*e*).—The switching transistor was unchanged. The *p–n* junction was the collector junction of type V30/10LP; the controlling transistor was a power output unit of type GT424 taken at random. This curve represents the best performance in the C state obtained during the investigation. The upper limit of current is well above the maker's rating for the switching transistor.

Curve (*f*).—This represents a switch of the opposite polarity; the switching transistor was now of type GT424, the controlling transistor of type V60/30P and the *p–n* junction was the collector junction of type V30/10LP. The upper limit of current was not measured but is well above the maker's rating.

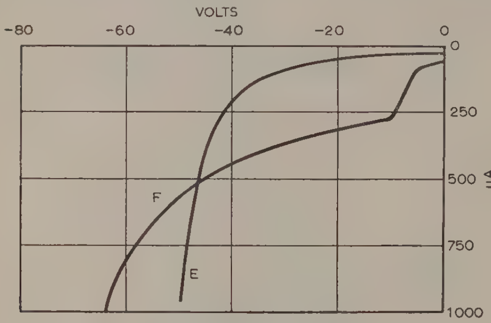


Fig. 14.—Reverse current in A state.

Fig. 14 shows the behaviour of the complete switch under reverse voltage in the A state; the two curves shown were measured on the same components as the corresponding curves of Fig. 13.

(9) CONCLUSIONS

The paper describes some elementary properties of controlle rectifiers using triode transistors and shows some design method The output currents and voltages of these circuits are determine by the properties of the transistors which are obtainable. A present it is possible to equal the current output of some *p–n–p–* diodes, but these latter have a roughly twofold advantage i inverse voltage. However, until there is a change in the relativ selling prices, which must all be regarded as tentative for sem conductor products, the complementary switch shows a perhap transitory but still considerable economy in first cost.

(10) ACKNOWLEDGMENTS

The work described was done in the Electrical Engineerin Department of the University of Birmingham, and the author anxious to record his gratitude to Prof. D. G. Tucker for makin available the facilities and resources of the Department.

(11) REFERENCES

(1) Shockley, W.: U.S. Patent No. 2655609.
(2) Deuitch, D. E., and Paz, H. J.: 'A Phase-Regulated Trar sistor Power Supply', *Transactions of the Institute of Radi Engineers*, 1957, **CT-4**, p. 279.
(3) Ebers, J. J., and Moll, J. L.: 'Large Signal Behaviour c Junction Transistors', *Proceedings of the Institute of Radi Engineers*, **42**, p. 1761.
(4) Ebers, J. J.: 'Four Terminal PNPN Transistors', *ibid.*, 195 **40**, p. 1361.
(5) Rissik, H.: 'Fundamental Theory of Arc Converters' (Chap man and Hall, 1939).
(6) Kinman, T. H., Carrick, G. A., Hibberd, R. G., an Blundell, A. J.: 'Germanium and Silicon Power Recti fiers', *Proceedings I.E.E.*, Paper No. 1936 U, Octobe 1955 (**103** A, p. 89).
(7) Gutzwiller, F. W.: 'The Current-Limiting Fuse as Fau Protection for Semiconductor Rectifiers', *Transactions c the American I.E.E.*, 1958, **77**, Part I, p. 751.

1.385.13.032.213

The Institution of Electrical Engineers
Paper No. 3308 E
Sept. 1960

ANODE LUMINESCENCE IN OXIDE-CATHODE RECEIVING VALVES

By H. N. DAGLISH, B.Sc., Ph.D., A.Inst.P., Graduate.

(*The paper was first received* 28*th February and in revised form* 10*th May*, 1960.)

SUMMARY

During life test, Post Office underwater-repeater valves and other ceiving-type valves exhibit a blue luminescent pattern on the inner ırfaces of the anode. The nature and some of the properties of the minescent material are examined, and it is concluded that the actual aterial is barium and/or strontium oxide transferred from the cathode ɔating during processing.

(1) INTRODUCTION

During the life testing of Post Office submarine-repeater alves[1, 2] it has long been observed that parallel bars of blue ıminescence are visible on the inner surfaces of the bright ickel anodes. Furthermore, the intensity of this luminescence as seemed to be at a maximum at the beginning of life, decreasing ith operating time. Similar effects have been observed in ommon receiving valves, and it is probable that the phenomenon a general one. The object of the present paper is to inquire nto the nature of the luminescent process and to see whether it as chemical implications which might affect the life of an oxide athode.

The phenomenon itself can be described in a few words. he blue luminescent area on the anode has the overall size nd shape of the oxide-cathode emitter, with the transverse bars ery obviously resulting from the shadowing action of the uppressor-grid wires. If the electron beam is deflected by neans of a magnet, the blue pattern on the anode moves, showing hat the whole surface is capable of luminescence wherever lectrons strike it.

It will be shown that this luminescence is due to a film of active naterial on the anode. The experimental work has been carried ɔut using coatings composed of single oxides of barium and trontium and of the equimolar double barium-strontium oxide prepared from the co-precipitated carbonates. Ammonium-precipitated carbonates were used throughout the work.

(2) THE LUMINESCENCE OF SPRAYED OXIDE COATINGS

An obvious source of active material is the cathode coating, which usually consists of double barium-strontium oxide. That such an oxide does exhibit luminescence when bombarded by electrons (i.e. cathodoluminescence) has been mentioned by a number of authors. Head[3] coated the screens of cathode-ray tubes with samples of many oxides, and obtained a yellow luminescence from strontium oxide but none from barium oxide. Aitchison,[4] among others, has referred to the luminescence from active cathode coatings as having a pale blue colour, fading as the temperature is increased to 700° K. Stout,[5] Noga and Nakamura,[6] and Gandy[7] have studied variations in the details of the band spectra of the luminescence during thermal activation of the oxides.

In order to have direct evidence of the colour to be obtained with the various oxides used in Post Office valves, special valves were constructed with two rectangular box-type cathodes, mounted facing each other and having independent heaters. In order to avoid any possible side-effects due to oxide from one coating contaminating the other, the two coatings were always of the same type of oxide. The coatings were applied by spraying as carbonates, which were converted thermally to oxide in the usual manner. The valves were gettered and sealed and then examined in total darkness. A potential of 150 volts was applied between the two cores, and the temperature of the negative core was increased until an electron current of between 1 and 10 mA flowed through the tube (i.e. current densities of about $2\frac{1}{2}$–25 mA/cm^2).

Coatings of the single barium and strontium oxides were examined, as well as the standard double barium-strontium oxide. All three types of coating behaved in a similar manner, and except where specially noted, all the colours described below were observed on specimens of each type. The standard core material for these experiments was 4% tungsten-nickel, but similar experiments were made with coatings sprayed on to substrates of platinum, molybdenum, pure nickel and O-nickel alloy. No real differences were observed between the luminescences of coatings on these various substrates.

Table 1

THE LUMINESCENCE OF BARIUM AND STRONTIUM OXIDES

Condition	Colour changes with activation increasing from left to right			
Hot	None	Light blue	Blue	
Cold	Yellow	Blue-green	Blue	Royal blue or purple

The colour of the cathodoluminescence was found to depend upon a number of factors, and was either yellow, green or blue. Spectroscopic examination of the light revealed a broad band spectrum with no detectable lines. A rough classification of the colours is made in Table 1. Because of the number of factors influencing the luminescence, no attempt has been made to define the parameters very precisely. However, the 'cold' regime extends up to a temperature of 600–700° K. The luminescent activation of the coating depends both upon its physical structure and upon the processing of the valve on the pump. Areas of coating much thicker than usual, because of overlapping of sprayed areas or the clumping of sprayed particles, tend to be much bluer than the surrounding areas. A high pressure (of carbon dioxide) during decomposition also increases the activity of the luminescent coating. The initial activation may subsequently be increased, at least temporarily, by the passage of current and decreased by overheating the coating. The royal blue and purple mentioned in Table 1 were quite uncommon, being observed only on a very small number of strontium and double-oxide coatings respectively. No similar colour was observed on any of the barium oxide coatings. Apart from

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.
Dr. Daglish is at the Post Office Research Station.

these small differences, the three types of coating behaved in a very similar manner.

The usual colour sequence observed in a new valve was as follows. The colour on first applying the anode voltage was usually yellow or—in a smaller number of valves—green-blue. Continued passage of current both activated and heated the anodic coating, and the luminescence diminished in intensity and changed in colour, becoming a rather pale and unsaturated blue. With an applied potential of 150 volts and a current of about 10 mA (i.e. about 25 mA/cm^2), the sequence of colours was complete within one or two minutes.

When the current was allowed to flow for several hours the blue colour became deeper and more saturated. Such continued passage of current through the coating eventually produced a visible stain on the white oxide. The colour of this stain was blue on the strontium oxide, grey-blue on the double oxide and grey-brown on the barium oxide. When the electron stream bombarded only part of the oxide coating the stain could be seen to correspond exactly to that part of the coating which was luminescent. When the electron stream was deflected magnetically, the luminescence of the stained area was seen to be of a much darker and bluer colour than that of the nearby oxide, which had been receiving comparatively few electrons. When the anodic coating was heated the stain disappeared. On reapplication of the anode voltage the luminescence was seen to have reverted to its original yellow colour. As current continued to flow the yellow turned blue and the stain slowly reappeared.

Eventually the glass envelope was cracked to allow air to enter. The stain on the oxide coating disappeared instantly, and was consequently identified as metallic barium or strontium. To determine whether this free metal had originated in the anodic or cathodic oxide layers, the experiments were repeated using a hot tungsten filament as the source of electrons. The stains which appeared were identical with those in the earlier experiments. It was also shown that, if the electron stream was deflected by fastening a small permanent magnet to the side of the envelope, the stain appeared only in those areas of oxide bombarded by electrons. It was therefore concluded that the stains were produced by the breakdown of the oxide when bombarded by electrons.

Experiments were carried out to detect the evolution of oxygen which would be expected to accompany the breakdown of the oxides. The experiments were made difficult by a

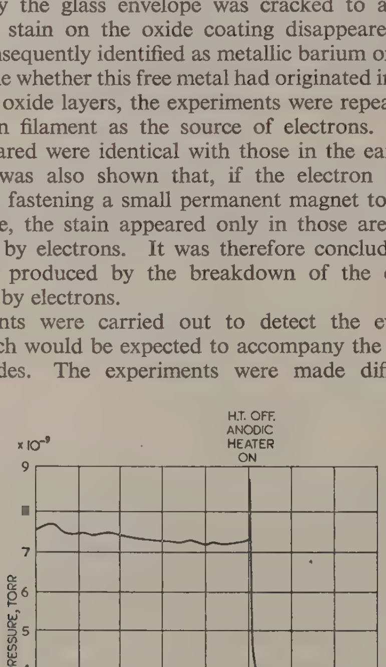


Fig. 1.—Gas evolution caused by electron bombardment of barium-strontium oxide.

tendency for oxygen ions to become trapped in the system. At normal operating temperatures the oxide matrix itself absorbs oxygen, and ions are also trapped in the electric field between the electrodes.

The experimental valves consisted of double-ended glass envelopes with a Bayard–Alpert ionization manometer in one end and two oxide-coated cathodes in the other. In order to reduce the absorption of gas by the oxide matrix, the cathode was maintained at a higher temperature than usual, about 1 100° K. The pressure was allowed to reach an equilibrium state while the anodic core was also maintained at this temperature by electron bombardment (approximately 22 mA of current at 160 volts). The anode voltage was then removed and the temperature of the anodic core was maintained constant by its internal heater. The equilibrium pressure in the valve was immediately observed to fall. An example of this is shown in Fig. 1. This rapid fall in pressure is attributed to a decrease in the rate of evolution of oxygen from the oxide on the anodic core. A similar fall in pressure was observed when the experiment was repeated with a similar valve in which the source of electrons was a heated tungsten filament.

(3) INTERPRETATION OF RESULTS WITH SPRAYED COATINGS

It is supposed that in the absence of free barium or strontium the oxides fluoresce under electron bombardment, with a yellow colour. The presence of a small quantity of free barium or strontium provides the additional activation centres which are necessary for the blue luminescence. This excess barium or strontium may already be present throughout the matrix as a result of the chemical processes occurring during the breakdown of the carbonate. Free metal may also be produced on the surface of the matrix by electron bombardment. The breakdown of barium and strontium oxides in this manner has also been described by Woods and Wright,[8] and Wargo and Shepherd.[9] The liberation of excess free metal in an oxide phosphor causes a reduction in the luminous efficiency of the phosphor—the so-called 'electron burn' described by Garlick.[10, 11] This was confirmed in the present experiments where the cathodoluminescence was always less bright in those areas of the matrix with a visible deposit of barium or strontium.

If the oxide layer on the anodic core is completely inactive with virtually no free barium, its electrical conductivity is low and its luminescence is yellow. In this condition the surface of the matrix may accumulate electrons faster than they can leak away through the matrix. Incoming electrons will then be repelled by the negative charge on the surface of the oxide and will be able to reach the anode only at an uncoated area. Such a phenomenon has been observed with oxide coatings which had been overheated to dissipate the free barium and which were bombarded with 70-volt electrons. The yellow area became covered with non-luminescent patches which spread to cover the whole coating except for the very thin layer at the edge of the sprayed coating. The effect disappeared when the applied voltage was increased. This increased the temperature and consequently the conductivity of the coating, as well as increasing the secondary electron emission from the oxide, and so allowed the surface charge to leak away. Similarly, the dark patches disappeared if the anodic core was heated merely by means of the insulated internal heater.

(4) THE LUMINESCENCE ON UNCOATED ANODES

Simple experimental assemblies were used to examine the luminescence on unsprayed anodes similar to those of pentodes; these assemblies consisted of two plates mounted on either side

f a standard cathode. The valves were pumped according to a andard processing schedule, based on that used for production peater valves, so that all the assemblies had the same treatment.

The anodes were heated by r.f. induction both before and fter thermal decomposition of the carbonates of the cathode pating. This decomposition was carried out using the same me and temperature sequence used for production pentodes. he tubes were gettered, sealed and based and then examined complete darkness. With the anode at +150 volts the mperature of the cathode was slowly increased until a current f about 5 mA was flowing (i.e. a current density of about mA/cm²). This was sufficient to make any luminescence isible without overheating the anode.

A number of these valves were examined in order to compare he cathodoluminescence with that of sprayed coatings of arium or strontium oxide. The luminescence on the unsprayed nodes took the form of a patch of colour, opposite the centre f the cathode. In most of the experiments the initial colour f this luminescence was a pale blue, although in some valves he initial colour was yellow, rapidly turning blue with the assage of current. After this initial colour had been observed he anodes were heated by r.f. induction to 1 100–1 200° K for 0 sec and then allowed to cool. When the valves were e-examined the cathodoluminescence was then either yellow or reen-blue. The passage of current rapidly restored the blue uminescence.

The behaviour of the luminescent material on the anodes was hus very similar to that of the sprayed oxide coatings. To confirm the presence of thermionically active material the electron current through one of the valves with an uncoated ungsten-nickel anode was increased so that the temperature of the anode was increased by bombardment to about 1 000° K. The polarity of the anode voltage was then suddenly reversed. For a few seconds as the anode cooled, sufficient electrons were emitted from it to produce a bright blue area of luminescence on the oxide-coated cathode.

It was thought that the nature of some of the other components in the assemblies might have some influence on the luminescence or on the transfer of material from the cathode. Some of the components were examined in the experiments summarized below:

Cathode Core.—When cathodes of tungsten-nickel and platinum were used, the cathodoluminescence produced on the anode was as described above. When O-nickel was used for the cathode core the luminescence was very similar, although initially somewhat more active. The luminescence did not exhibit any yellow phase.

Coating Density.—There was no difference in the colour of luminescence produced on the anode by cathodes of various densities, but coatings of an open texture with low density appeared to transfer more material to the anode than did smooth high-density coatings, which took much longer to decompose from carbonate to oxide.

Binder.—Normal cathodes contain an organic binder to hold the particles of carbonate together. While this binder should decompose or volatilize during the baking of the valve during pumping, it is possible that residual traces might be responsible for some of the observed cathodoluminescence. Some cathodes were therefore sprayed without binder, using only a suspension of barium carbonate in methyl alcohol. Although the coating so produced was very thin, the cathodes behaved in a manner very similar to standard cathodes. When bombarded, the initial yellow luminescence turned successively green and then blue. When these cathodes (on both tungsten-nickel and platinum cores) were used between plain molybdenum anodes, the usual pale blue luminescence was produced on these anodes.

Anode Materials.—A number of different pure metals were used as anodes,* and the blue luminescence appeared to be identical on all of them when the valves were directly compared. Several nickel alloys were also examined. Whereas the tungsten-nickel behaved in the same manner as the pure metals, the behaviour of the more active nickel alloys depended on the processing of the anode before assembling into the valve. Simply degreased, the anodes behaved as described above, the appearance of the luminescence being just as on the pure metals. Heating the anodes to red heat deactivated the luminescent layer on all of the metals mentioned, except for titanium and a nickel alloy containing about 0·2% of silicon. After heating the anode the colour of these two remained blue, whereas that on all the others was initially yellow or green-blue, becoming blue again only after the passage current.

The behaviour of the active nickel alloys was quite different if the anode processing had included any hydrogen stoving. Anodes treated in this way became luminescent over the whole surface, the colour being initially pale blue and becoming purple-red after heating. This behaviour is attributed to the formation of luminescent oxides of magnesium and silicon on the surfaces of the anode. Experiments with oxidized pure nickel failed to show any luminescence which could be attributed to the nickel oxide.

From these experimental data, it is concluded that the luminescence on the anodes of simple diodes is due to the presence on the anodes of active cathode-type material. It is reasonable to assume that the luminescence on the anodes of pentode repeater valves has the same origin, although the more elaborate pentode pumping and ageing schedule will affect the initial activity of the layer of oxide on the anode.

The layer of oxide must have been transferred during the processing of the valve. The cathode is very hot during part of the processing and some oxide will be evaporated. However, it seems probable that much of the transfer takes place during the decomposition of the cathode coating when the rate of evolution of carbon dioxide is very high, possibly by very fine particles being carried directly from the surface of the oxide to the anode. The amount of material transferred depends upon the porosity of the sprayed coating, the temperature of the cathode, and the rate of evolution of carbon dioxide. A transfer mechanism depending upon the evolution of carbon dioxide would explain why subsequent attempts to evaporate more active material had comparatively little effect. When a processed cathode was taken through the complete time and temperature sequence as used during the decomposition of the carbonate, there was little increase in the intensity of the luminescence on the anode.

(5) CONCLUSIONS

It has been shown that the blue luminescence on receiving-valve anodes indicates the presence of a film of barium and/or strontium oxide, activated by small amounts of excess free barium or strontium metal. The luminescence itself can have no deleterious effects on the valve, but the presence of the oxide on the anode might have some such effects.

The layer of oxide will be slowly decomposed by the continuous bombardment of electrons. Although some recombination will take place, some barium, strontium and oxygen must be released into the valve. This slow evolution has been shown to have a negligible effect on the behaviour of receiving-type pentodes, by comparing the life test results for standard valves with those for experimental valves whose anodes were mechanically screened during decomposition and activation of the cathode coating.

* Platinum, molybdenum, tantalum, titanium and pure nickel.

The presence of the luminescent oxide layer on the anodes of receiving-type valves is therefore not considered dangerous to normal operation of the valves.

(6) ACKNOWLEDGMENTS

Acknowledgment is made to the Engineer-in-Chief of the Post Office for permission to make use of the information contained in the paper.

(7) REFERENCES

(1) McNally, S. O., Metson, G. H., Veazie, E. A., and Holmes, M. F.: 'Electron Tubes for the Transatlantic Cable System', *Proceedings I.E.E.*, Paper No. 2279 R, January, 1957 (Suppl. 4, **104** B), p. 60.
(2) Holmes, M. F., and Reynolds, F. H.: 'Post Office Valves for Deep-Water Submarine Telephone Repeaters', *ibid.*, Paper No. 3163 E, March, 1960 (**107** B, p. 165).
(3) Head, R. B.: 'The Luminescence of Simple Oxide Phosphors', *Electronic Engineering*, 1948, **20**, p. 219.
(4) Aitchison, R. E.: 'The Luminescence of Barium Strontiu Oxide', *Nature*, 1949, **164**, p. 1088.
(5) Stout, V. L.: 'Luminescence and Thermionic Emission Barium Oxide', *Physical Review*, 1953, **89**, p. 310.
(6) Noga, K., and Nakamura, K.: 'Cathodoluminescen Study of (Ba, Sr)O', *Vide*, 1954, **9**, 113.
(7) Gandy, H.: 'Cathodoluminescence of Strontium Oxid Barium-Strontium Oxide and Magnesium Oxide', *Physic Review*, 1958, **111**, p. 764.
(8) Woods, J., and Wright, D. A.: 'The Decomposition Thin Films on Bombardment with Slow Electron *Proceedings of the Physical Society*, 1953, **55 B**, p. 107
(9) Wargo, P., and Shepherd, W. G.: 'Electron-Bombardmen Induced Dissociation of Alkaline Earth Oxides', *Physic Review*, 1957, **106**, p. 694.
(10) Garlick, G. F. J.: 'Cathodoluminescence', *Advances Electronics*, 1950, **2**, p. 151 (see Section 5*a*).
(11) Garlick, G. F. J.: 'Luminescent Materials' (Oxfo University Press, 1949), p. 187.

MONOGRAPHS PUBLISHED INDIVIDUALLY

Summaries are given below of monographs which have been published individually, price 2s. each (post free). Applications, quoting the serial numbers as well as the authors' names, and accompanied by a remittance, should be addressed to the Secretary. For convenience, books of five vouchers, price 10s., can be supplied.

The Magnetic Excitation inside a Cylindrical Thin-Film Ferromagnet. Monograph No. 396 M.

T. H. O'Dell, B.Sc.

An expression for the magnetic excitation inside a cylindrical thin-film ferromagnet is derived, and a Table of computed values is given The results are considered to be relevant to work on thin ferromagnetic films for digital-storage applications.

The Conductivity of Oxide Cathodes: Part 9—Thermo-electric Power. Monograph No. 397 E.

G. H. Metson, M.C., D.Sc., Ph.D., M.Sc., B.Sc.(Eng.), and M. F. Holmes, B.Sc.

The thermo-electric power of an oxide cathode has proved to be a complex function of temperature, dependent on the dual nature of oxide-cathode conductivity. The authors show that an apparently complex form of behaviour is, in fact, the result of the superposition of two quite simple phenomena. Two parallel-acting thermo-electric power functions are involved, and each of these is invariant with temperature and temperature gradient. The two functions are physically separated and each is measured over an appropriate temperature range. The larger function, 2·0 — 3·0 mV/degC, is associated with the vacuum movement of electrons through the hollow pores of the oxide matrix; the smaller one, 0·5 mV/degC, occurs in the chains of contiguous solid particles of the matrix. It is concluded that a temperature gradient through an oxide matrix leads to a continuous circulation of current, vacuum-wise in one direction and solid-wise in the other.

Forced Oscillation in an Oscillator with Two Degrees of Freedom. Monograph No. 398 E.

B. R. Nag, M.Sc.Tech.

Resonance curves of an oscillator with two degrees of freedom and the criteria of stability of the synchronized state are determined theoretically by van der Pol's method of approximation and verified experimentally by an electronic differential analyser. An experimental method for obtaining the phase trajectories, and characteristics of oscillations outside the zone of synchronization, are also described.

Matrix Analysis of Constrained Networks. Monograph No. 399 E.

A. Nathan, M.S.

It is shown how networks with imposed unilateral constraints, i transmittances and voltage sources, can conveniently be analysed matrix methods. Their admittance matrix is equal to the product a submatrix of the admittance matrix before the application of t constraints and a transmittance matrix which describes the contraint Constraints lower the rank of the original admittance matrix, th making the method quite attractive in practice. Applications includ are a computing network, difference amplifiers, a d.c. amplifier a signal-flow graphs.

Limitations of Realizable Response Shapes for Certain Wide-Ba Band-Pass Amplifier Circuits. Monograph No. 400 E.

R. A. Woodrow, B.Sc.

A procedure is described for the design of wide-band band-pa amplifiers using either stagger-tuned stages or a chain of feedba pairs to realize Chebsy̆hev or Butterworth response shapes.

A graphical test has been developed by which it is possible to veri at the outset, that a given design specification is a realizable o This reveals certain fundamental limitations which are imposed up the response shapes by the choice of these circuit configurations.

An example demonstrating the application of the procedure to design problem has been included to demonstrate that the procedu for a Chebyshev response is no more complicated than for a Butte worth response, neither is any additional complication introduced increasing the number of singularities in the desired response sha beyond the first pair, since each pair of singularities are separate realized.

Application of the Theory of Orthogonal Polynomials in Two Variabl to a Multi-gain Equivalent Linearization Problem. Monograph No. 401 M.

J. L. Brown, Jr., Ph.D.

It is shown that the multi-gain representation for a single-valu non-linearity with multiple inputs as developed by Somerville a Atherton may be regarded as an approximation problem involvi orthogonal polynomials in two variables.

Consider two random processes, $x(t)$ and $y(t)$, possibly correlate with a given second-order joint probability density, $p(x, y)$. If t input to a specified, zero-memory non-linear device having input/outp characteristic $v_0(t) = f[v_i(t)]$ is $x(t) + y(t)$, then the relevant po nomials satisfy orthonormality conditions over the x-y plane wi respect to $p(x, y)$ as weighting function. An inherent minimu property of these polynomials then allows the equivalent gains to determined directly in terms of the expansion coefficients of $f(x +$ with respect to the polynomials. When x and y are uncorrelated, t gains reduce to the values previously obtained by Somerville a Atherton.

PROCEEDINGS OF THE INSTITUTION OF ELECTRICAL ENGINEERS

Part B. ELECTRONIC AND COMMUNICATION ENGINEERING (INCLUDING RADIO ENGINEERING), SEPTEMBER 1960

CONTENTS

Declaration on Fair Copying.—Within the terms of the Royal Society's Declaration on Fair Copying, to which The Institution subscribes, material may be copied from issues of the *Proceedings* (prior to 1949, the *Journal*) *which are out of print and from which reprints are not available.* The terms of the Declaration and particulars of a Photoprint Service afforded by the Science Museum Library, London, are published in the *Journal* from time to time.

Bibliographical References.—It is requested that bibliographical reference to an Institution paper should always include the serial number of the paper and the month and year of publication, which will be found at the top right-hand corner of the first page of the paper. This information should precede the reference to the Volume and Part. *Example.*—SMITH, J.: 'Reflections from the Ionosphere', *Proceedings I.E.E.*, Paper No. 4001 R, December, 1954 (**102** B, p. 1234).

Published by The Institution, Savoy Place, London, W.C.2. *Telephone* COVent Garden 1871. *Telegrams:* 'Voltampere, Phone, London.'
Printed by Unwin Brothers Limited Woking and London.